Abundance and diversity of marine microbial eukaryotes
(Abundacia y diversidad de eucariotas microbianos marinos)

Massimo Ciro Pernice

Tesis Doctoral presentada por Massimo Ciro Pernice para obtener el grado de Doctor por la Universidad de las Palmas de Gran Canaria, Departamento de Biología, Programa en Oceanografía
(Biennio 2008-2010)

Director: Ramon Massana i Molera

Universidad de las Palmas de Gran Canaria
Institut de Ciències del Mar (ICM-CSIC)

El Doctorando
Massimo Ciro Pernice

El director
Ramon Massana i Molera

En Barcelona, a de de 2014

A mi Familia filogénetica y
a mi Comunidad biogeográfica

Este libro en su conjunto no es más que un borrador; mejor dicho, el borrador de un borrador.
¡Oh, tiempo, fuerza, dinero y paciencia!

(Herman Melville, Moby Dick)

Contents

Summary/Resumen/Resum 13

Introduction 19

Aims and outline 31

Chapter 1 Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean 37

Chapter 2 General patterns of diversity in major marine microeukaryote lineages 63

Chapter 3 Global abundance of planktonic heterotrophic protists in the deep ocean 95

Chapter 4 Diversity of marine microeukaryotes in the global deep ocean 121

Synthesis of results and general discussion 151

Resumen en español 163

General references 213

Agradecimientos 223

Summary

Microeukaryotes are important ecological players in any kind of ecosystem, most notably in the ocean, and it is therefore essential to collect information about their abundance and diversity. To achieve this general goal this thesis was structured in two parts. The first part represents an effort to define our "diversity unit" from studies based on the well-known cloning and Sanger sequencing approach. Basically, we wanted to establish a solid baseline for the second part of the thesis. We started with data from one cruise (Chapter 1) and then continued with the analysis of the complete dataset of 18S rDNA sequences available at that time (Chapter 2). From this analysis we found that the V4 region of the 18S rDNA was a good proxy of the variability of the entire gene. We also determined that the maximal genetic distance for sequences belonging to a same class was 0.25. Once defined this framework, it was used in the second part of the thesis for studying deep ocean microeukaryotes. Thanks to the Malaspina 2010 expedition, we had a comprehensive set of deep samples with associated abiotic and biotic parameters from all over the world. We found that the microeukaryotes abundance averaged 54 cells mL^{-1} in the mesopelagic layer and 14 cells mL^{-1} in the bathypelagic layer, and its variability was explained by depth, prokaryotes abundance and oxygen concentration (Chapter 3). Finally, the diversity of deep microeukaryotes was determined by pyrosequencing and metagenomic tags (Chapter 4). The bathypelagic community was mainly composed by Collodaria, Chrysophyceae, MALV-II and Basidiomycota. However, the relative abundance of these classes varies a lot among samples. The variability in community composition between samples was well explained by the water mass they belong and by the abundance ratio between prokaryotes and microeukaryotes.

Resumen

Los Microeucariotas son actores ecológicos importantes en cualquier tipo de ecosistema, sobre todo en el océano, por lo que es esencial recopilar información acerca de su abundancia y diversidad. Para lograr este objetivo general esta tesis se ha estructurado en dos partes. La primera parte representa un esfuerzo para definir nuestra "unidad de diversidad", empezando por estudios basados en la clonación molecular y la secuenciación de Sanger. Básicamente, queríamos establecer una base sólida para la segunda parte de la tesis. Empezamos con los datos de una campaña (Capítulo 1) y luego seguimos con el análisis del conjunto completo de datos de secuencias de 18S ADNr disponibles en ese momento (Capítulo 2). A partir de este análisis, se encontró que la región V4 del 18S ADNr es un buen indicador de la variabilidad de todo el gen. También se determinó que la distancia genética máxima para las secuencias que pertenecen a una misma clase es de 0.25. Una vez definido este marco, fue utilizado en la segunda parte de la tesis para estudiar los microeucariotas del oceáno profundo. Gracias a la expedición Malaspina 2010, disponíamos de un amplio conjunto de muestras de profundidad de todo el mundo y de sus parámetros abióticos y bióticos asociados. Se encontró que la abundancia de microeucariotas promedio era de 54 células mL^{-1} en la capa mesopelágica y de 14 células ml^{-1} en la capa batipelágica. Su variabilidad se explicaba por la profundidad, la abundancia de procariotas y la concentración de oxígeno (Capítulo 3). Por último, la diversidad de microeucariotas profundos se determinó mediante pirosecuenciación y secuencias de metagenómica (Capítulo 4). La comunidad batipelágica estaba compuesta principalmente por Collodaria, Chrysophyceae, MALV-II y Basidiomycota. Sin embargo, la abundancia relativa de estas clases varía mucho entre las muestras. La variabilidad en la composición de la comunidad entre las muestras se explicaba bien teniendo en cuenta la masa de agua a la que pertenecían y el ratio de abundancia entre procariotas y microeucariotas .

Resum

Els microeucariotes són importants actors ecològics en qualsevol tipus d'ecosistema, sobretot en l'oceà, per la qual cosa és essencial recopilar informació sobre la seva abundància i diversitat. Per aconseguir aquest objectiu general aquesta tesi s'ha estructurat en dues parts. La primera part representa un esforç per definir la nostra "unitat de diversitat" començant per estudis basats en la clonació molecular i seqüenciació de Sanger. Bàsicament, volíem establir una base sòlida per a la segona part de la tesi. Hem començat amb les dades d'un creuer (Capítol 1) i després hem seguit amb l'anàlisi del conjunt complet de dades de seqüències de 18S ADNr disponibles en aquest moment (Capítol 2). A partir d'aquesta anàlisi, es va trobar que la regió V4 del 18S ADNr és un bon indicador de la variabilitat de tot el gen. També es va determinar que la distància genètica màxima per a les seqüències que pertanyen a una mateixa classe és de 0.25. Un cop definit aquest marc, va ser utilitzat en la segona part de la tesi per estudiar microeucariotes de l'oceà profund. Gràcies a l'expedició Malaspina 2010, teníem un ampli conjunt de mostres profundes associades a paràmetres abiòtics i biòtics d'arreu del món. Es va trobar que l'abundància mitjana de microeucariotes era de 54 cèl·lules mL^{-1} a la capa mesopelàgica i 14 cèl·lules mL^{-1} a la capa batipelàgica, i la seva variabilitat s'explicava per la profunditat, l'abundància de procariotes i la concentració d'oxigen (Capítol 3). Finalment, la diversitat de microeucariotes de l'oceà profund es va determinar mitjançant piroseqüenciació i seqüències de metagenòmica (Capítol 4). La comunitat batipelàgica estava composta principalment per Collodaria, Chrysophyceae, MALV-II i Basidiomycota. No obstant això, l'abundància relativa d'aquestes classes varia molt entre les mostres. La variabilitat en la composició de la comunitat entre les mostres queda ben explicada per la massa d'aigua de pertinença i per la ràtio d'abundància entre procariotes i microeucariotes.

Introduction

Marine protist research: a brief history

There is an extremely broad diversity of organisms that fall within the term “protist”. Generally speaking, protists are eukaryotic microorganisms. In fact, this term does not have a real evolutionary meaning since it includes all eukaryotes that are not animals, plants or fungi. The first registered observation of a protist was done by Leeuwenhoek in 1674, but the term was coined and popularized in 1866 by Haeckel (famous for his detailed illustrations of these organisms, Figure 1), and at the beginning this term also comprised the prokaryotes. *Protista* was considered as a new life kingdom, together with animals and plants, apparently less important and with few categories. Today, thanks to several studies that began in the second half of the XX century, we know that animals and plants are two little leafs in the phylogenetic tree of eukaryotes, which is mostly formed by unicellular forms of life (Figure 2).

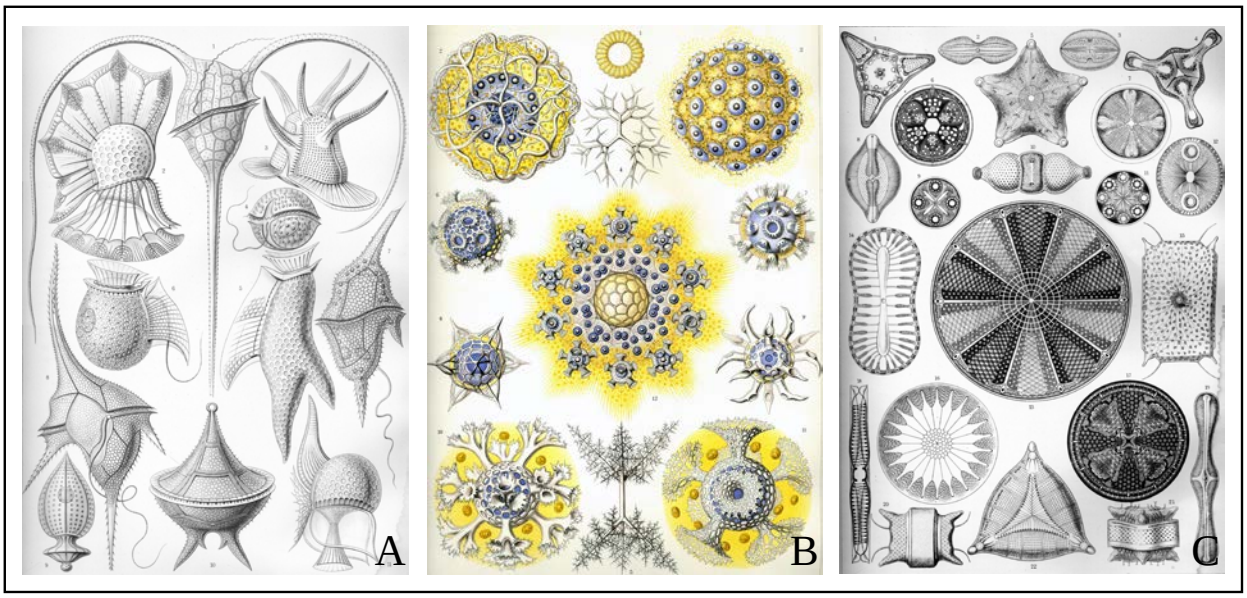


Figure 1. Morphological diversity in Alveolata (A), Rhizaria (B) and Stramenopiles (C). Drawings by Ernst Haeckel, 1904.

Methodological improvements made possible to skip from the study of the visible to the discovery of the invisible. Thus, several techniques are fundamental for the study of microorganisms. A classical method is the observation and counting of protist cells with epifluorescence microscopy, which implies the utilization of cellular stains, as for example DAPI (4’,6-diamidino-2-phenylindole) that binds to the DNA of the cells (Porter and Feig 1980). A more focused technique is Fluorescent In Situ Hybridization (FISH, Pernthaler *et al.* 2003, Massana *et al.* 2006) that targets specific cells using taxon-specific oligonucleotide probes, allows collecting information about the abundance and the global diversity of marine protists (Morgan-Smith *et al.* 2011 and 2013), and can also be used in grazing experiments (Fu *et al.* 2003, Jezbera *et al.* 2005, Massana *et al.* 2009). Another method to quantify microbial abundance is flow-cytometry (Zubkov *et al.* 2006

and 2007; Christaki *et al.* 2011). This technique is really useful with a large amount of samples, but still needs to be improved and optimized for some functional groups like heterotrophic micro-eukaryotes. Despite the possibility to identify some cell types, epifluorescence microscopy is not enough accurate for diversity studies. Indeed, morphology is useless to identify picosized cells (and many nanosized cells), and often does not allow to go deep inside the taxonomy at species level.

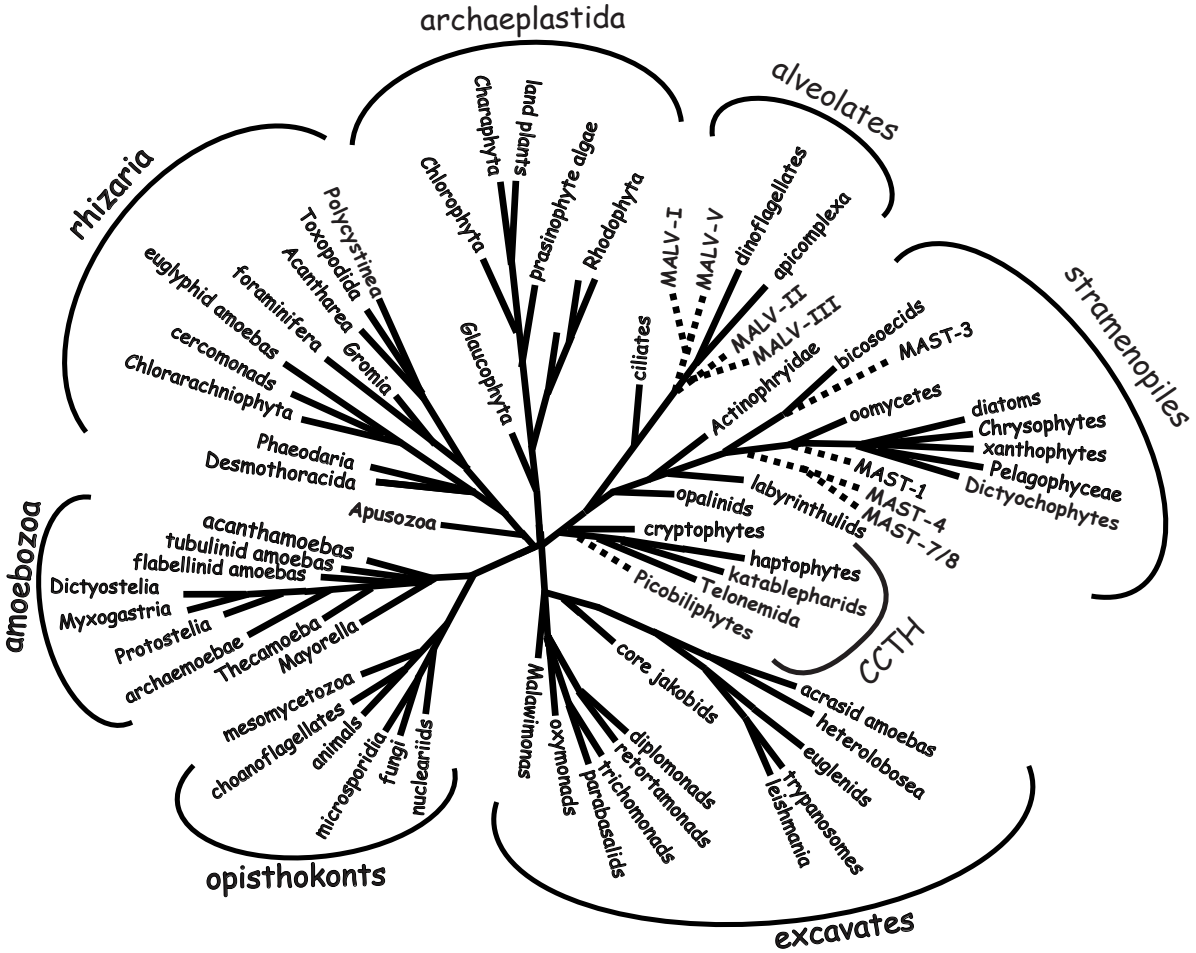


Figure 2. Eukaryotic tree of life, showing the consensus phylogeny of the major eukaryotic groups based on molecular and ultrastructural data (adapted from Baldauf 2003). Dotted lines indicate positions of major lineages known primarly from culture-independent molecular surveys.

With the development of molecular methods the study of microbial diversity improved exponentially. Pioneer protist studies in this direction were those of Díez *et al.* 2001, López-García *et al.* 2001 and Moon-van der Staay *et al.* 2001 targeting picoeukaryotes. These investigations had to face a general problem: to estimate the diversity is fundamental to identify groups of similar organisms that are named "species" in the classical taxonomy. Many different species concepts have been applied to microorganisms in general and protists in particular (Roselló-Mora and Amann 2001, Schlegel and Meisterefeld 2003). The most pragmatic concept proposes that a species is a

"group of organisms that share similar morphological characteristics". The biological concept, perhaps the most useful in animals and plants, defines a species as "a group of organisms capable of interbreeding and producing fertile offspring". Although most protist cell divisions are asexual, sexual reproduction is also known to be present in protists (Amato *et al.* 2007), but there is little information about how spread it is throughout the different protistan groups and how frequent it occurs. Of course, life-cycle studies of protist species are necessary to find out the incidence of asexual-sexual divisions. At present the principal limitation of these life-cycle studies is that only a few protist species are cultured and well-characterized, and even some groups completely lack a cultured representative. So, it is not practical to invoke the biological species concept for studying protist diversity.

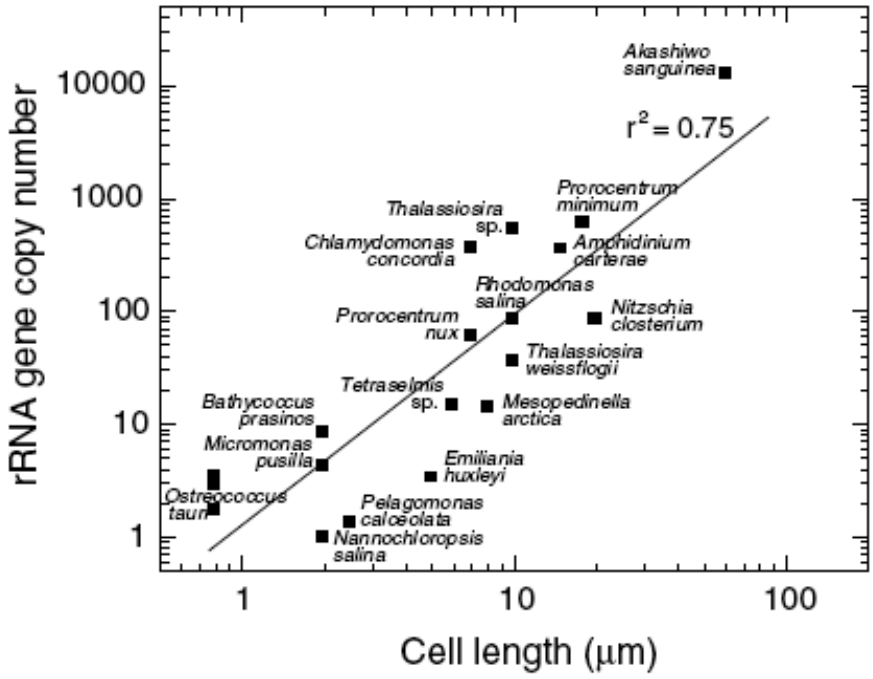


Figure 3. Correlation between cell size and rDNA copy number in different protist species. Taken from Zhu *et al.* (2005).

Luckily for microbiologists, during the 70s of the past century Carl Woese came out with the idea that it was possible to identify and organize all life forms by comparing their DNA sequences (Woese and Fox 1977). This operation needed basically two steps: alignment of DNA sequences of the same gene and measurement of their genetic distances. The preferred target gene for this approach since the beginning was the ribosomal RNA gene (rDNA). This gene is present in all organisms and it is conservative enough to be used in phylogeny among any life form. Molecular taxonomy has several advantages: it can be applied to a wide range of taxa, to all life stages and to

Table 1. Intragenomic 18S rDNA gene variability (SSU).

Reference	Species	Sim	Origin
Rooney *et al.* (2004)	*Cryptosporidium parvum*	92.1	Genome
	Plasmodium Falciparum	89.5	Genome
	Plasmodium berghei	92.0	GeneBank
Alverson *et al.* (2005)	*Skeletonema grethae*	99.2	Strain
	Skeletonema japonicum	99.4	Strain
	Skeletonema menzeleii	99.4	Strain
	Skeletonema pseudocostatum	99.5	Strain
	Skeletonema subsalsum	99.5	Strain
Simon *et al.* (2008)	*Phoma exigua*	99.5	Strain
	Mycospharella punctiformis	99.6	Strain
	Teratospheria microspora	99.6	Strain
	Davidiella tassiana	99.4	Strain
	Aspergillus nidulans	99.6	Strain
Gong *et al.* (2013)	*Tintinnopsis sp.*	99.1	Individual
	Pseudotontonia sp.	99.3	Individual
	Strombidium sp.	99.7	Individual
	Vorticella sp.	99.1	Individual

Values of intragenomic genetic similarity in different microeukaryotes species. Low values of similarity in *Plasmodium spp*. and *Cryptosporidium parvum* are explained by the effective presence of different ribosomal forms activated in different hosts of these parasites.

the large number of data that are typical of most ecological studies (Caron *et al.* 2009). The rDNA is very useful but it is not a perfect target, since it is typically a multi-copy gene, particularly in eukaryotes. In algal strains, the copy number ranges from 1 to 10,000 (Zhu *et al.* 2005) implying that relative gene abundance can deviate strongly from relative cell abundance. The copy number is proportional to cell-size and genome size (Figure 3) so the chances of great variations is lower for pico and nano sized cells. Moreover, it is possible that these copies have some variability at intragenomic level. The risk of intragenomic variability is that we could detect two or more different sequences when there is only one organism. Again, in most cases this intragenomic variability is very low (Table 1).

The important innovation of the molecular techniques was the possibility of a more realistic study of marine microbial diversity, particularly concerning nano- and picosized plankton. The seminal approach was the construction of clone libraries of 18S rDNA genes, which were amplified from environmental genomic DNA by a polymerase chain reaction (Saiki *et al.* 1985) step. Typically, between 100 and 500 sequences were obtained per clone library. These ribosomal sequences became the basis for a new molecular taxonomy; in fact a new “species concept” more pragmatic than the biological or morphological criteria appears: that related to operational taxonomic units (OTUs). Following this criterion, sequences are grouped in countable units that have a certain degree of genetic divergence, chosen by the researcher, in an operation commonly known as clustering. Is important to highlight that the way that sequences are clustered in OTUs is a crucial step

that determines our vision of the diversity in marine samples.

Despite some limitations, the rDNA gene (and particularly the 18S rDNA) is still the best compromise to study protist diversity and has been chosen as target in the emergent high-throughput sequencing (HTS) technology (454 and Illumina), which has evolved so fast that the initial definition of "Next-generation sequencing (NGS)" has become obsolete in less than five years. The number of sequences collected in HTS is several orders of magnitude higher than the one obtained in clone libraries, and then a new problem appears related with the management of these huge amounts of data. However there is a great enthusiasm about the possibility of "sequencing the ocean" (Venter *et al.* 2004) and many scientists are working in the optimization of the methods and improve the confidence of the approach (Kunin *et al.* 2009; Quince *et al.* 2009). High-throughput sequencing gives the possibility to go deep inside in diversity studies. It is important to remember that, when combined with PCR amplicons, HTS is subjected to the same PCR biases (Wintzingerode *et al.* 1997). Nowadays metagenomics, despite having as a principal goal the study of metabolic functions more than species' diversity, is a viable alternative for the collection of 18S rDNA sequences from natural microbial assemblages (Logares *et al.* 2013). The use of metagenomic techniques is independent of the PCR step, so eliminates this source of errors. To better understand the inner characteristics of protist diversity, all the different approaches described have been used in this thesis.

General taxonomy of microeukaryotes

The high-rank taxonomy of eukaryotes at the present time is a continuous matter of debate in the scientific world (Burki *et al*, 2008). Thanks to the combination of microscopy and molecular biology it is possible to identify most taxa, but the real challenge is to understand how these taxa are related among them. In this thesis there is a mix of classical morphological taxa (better defined thanks to molecular tools) and new ribogroups, which are formed by sequences that cluster together in a tree and branch outside the well known groups. So, the new ribogroups are inserted among the classical taxa well defined by morphology, which are used as the backbone of the eukaryotic tree of life. The general taxonomy reference for most morphological groups follows the classification of Adl *et al.* (2012).

Among the entire eukaryotic tree of life (Figure 2) four protists supergroups deserve to be mentioned due to their importance in molecular surveys of protists in the marine environment: Alveolata, Rhizaria, Stramenopiles and CCTH. Alveolata, often the most abundant supergroup, comprises two of the most studied classical classes: dinoflagellate and ciliates. Rhizarian are composed by Radiolaria, which can have solitary or colonial lifestyles and are characterized by complex structures, Cercozoa (also known as Filosa) and Foraminifera, which prefer living in sediments. Stramenopiles encompass phototrophic groups like diatoms, as well as heterotrophic groups such as bicosoecids. The CCTH is a recently proposed supergroup that includes Cryptophyta, Centroheliozoa, Telonemia, and Haptophyta (Burki *et al.* 2009), but more recent phylogenies raises some doubts about its monophyly (Hampl *et al.* 2009, Baurain *et al.* 2010, Burki *et al.* 2012). In the last decade, and thanks to molecular surveys, a "forgotten" group has risen in importance in the oceanic ecosystem, the Fungi (Bass *et al* 2007, Richards *et al.* 2012). For traditional reasons, Fungi were generally studied by botanists and not protistologists. However, since many fungal species are unicellular, they perfectly fit within the microeukaryotes targeted in marine studies. Fungi have been previously found in seawaters, including the deep ocean ecosystem (Bass *et al.* 2007, Jebaraj *et al.* 2009, Edgcomb *et al.* 2011, Richards *et al.* 2012). Initially, fungal sequences were disregarded in protist surveys, like metazoan sequences, but now they are appreciated and kept. In fact, it does not make sense to exclude such important marine players.

Recently defined ribogroups represent a great part of retrieved sequences in marine molecular surveys. In fact the majority of the sequences of this thesis belong to Marine Alveolates (MALV) that were already detected in the first molecular surveys of deep marine waters (López-Garcia *et al.* 2001) and better defined later (Groissillier *et al.* 2006). Other important ribogroups are the Marine Stramenopiles (MAST), defined in 2004 by Massana *et al.* and the Picozoa (known before

as Picobiliphyta) that were first identified by environmental sequences (Not *et al.* 2007a) and later cultivated (Seenivasan *et al.* 2013). In Rhizaria there are three ribogroups, RAD-A, RAD-B, and RAD-C, the second encompassing the former morphological group of Sticholonche previously known as Taxopodia. All these ribogroups are now widely found and recognized, therefore entering de facto in "practical" taxonomical schemes.

Trophic roles and participation in biogeochemical cycles

In one millilitre of epipelagic seawater there are about 1000-10,000 cells of microeukaryotes. It is difficult to identify the ecological function of each different taxa but it is clear that together they play important roles in biogeochemical cycles, both as autotrophs and heterotrophs. It is worth to remember that phytoplankton, today considered as the sum of phototrophic bacteria and phototrophic protists, produces 70% of the total oxygen of the planet (Epstein *et al.* 1993) and makes life on Earth possible. Generally unicellular organisms are connected in a complex size-based trophic webs. The microbial loop, proposed by Azam *et al.* in 1983, constitutes an interesting hint of this web. Through this loop, the dissolved organic matter is consumed by prokaryotes and arrives to upper trophic levels thanks to the fact that phagotrophic protists feed on prokaryotes and are then fed by larger zooplankters (Figure 4 and Figure 5b). Prokaryotes could be considered as the biochemical machines that drive the principal biogeochemical cycles (carbon, nitrogen, sulfur), but at the end who controls the velocity of these metabolic reactions are the bacterivorous protists, probably together with viruses (Boras *et al.* 2010).

Phagotrophy, the ingestion of food particles through engulfment of the cell membrane, is widespread among protist taxa. Both in cultured species and in environmental samples (mostly in epipelagic waters) is a quite well studied process. Important grazer classes are for example the ciliates, which are the predators in the classical food web, and the chrysophytes or the MAST-4 ribogroup, perhaps the most important bacterivorous in the marine microbial loop (Massana 2011). Nevertheless, phagotrophy is not the only form of heterotrophy in the environment. Several species belonging to Fungi (Richards *et al.* 2012), Excavata (von Der Heyden *et al.* 2004), Chrysophyceae (Holen and Boraas 1996, Sanders *et al.* 2001) or Labyrinthulidae (Raghukamar *et al.* 2001) survive by osmotrophy, which is the uptake of dissolved organic compound by osmosis. However the prevalence of this phenomenon is still not well understood. In addition, there are several examples of groups that survive in the ocean thanks to parasitic interactions with a varied array of marine hosts. These marine parasites include the MALV-I and -II ribotypes, the most abundant groups in terms of sequences retrieved (Siano *et al.* 2010), pirsonids (Schnepf *et al.*

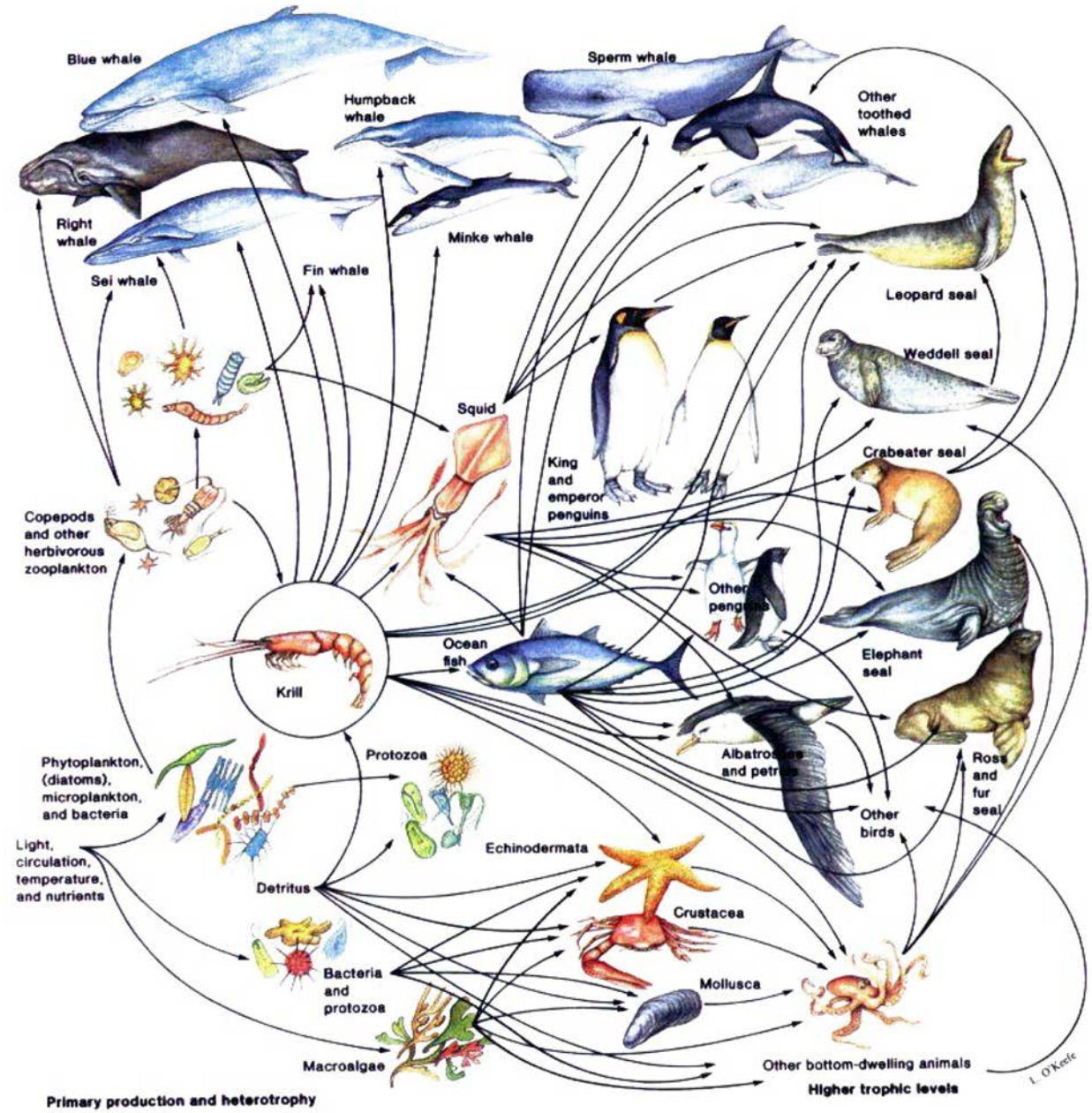


Figure 4. A complete marine food web, indicating a large array of species and their interactions. The microbial food-web is shown on the left, and highligths the trophic connections between microbiota and macrobiota. Drawing by O'Keefe L.

1990) and several fungal species (Richards *et al.* 2012). The environmental study of parasitism is quite hard and despite a few documented cases (Chambouvet *et al.* 2008), the information about the magnitude of the global phenomenon is, as for osmotrophy, poorly understood. Apart from these strict trophic divisions, it is important to remember that the unicellular world seems to favor a sort of plasticity in the trophic style, and mixotrophy, the combination of autotrophy and heterotrophy in the same organism, appears as a common behaviour between protists taxa (Sanders *et al.* 1991, Jones 2000, Zubkov *et al.* 2008).

The deep ocean: a peculiar habitat

The ability of adaptation of microeukaryotes is evident by the fact that they are widespread in the planet, including all sorts of extreme environments. In the ocean, we know that they are present in the entire water column (Not *et al.* 2007b). However, for obvious reasons, surface communities are much better studied than deeper ones. In fact, the functioning of the deep ocean ecosystem is far from being completely clear. Traditionally the deep dark ocean is divided in three zones, the mesopelagic (200-1000 m), the bathypelagic (1000-4000 m) and the abyssopelagic (more than 4000 m). The mesopelagic layer, where often resides the thermocline, appears to be more influenced than the other two deeper layers by the epipelagic system (0-200 m). Indeed, a large fraction of the organic carbon fixed by photosynthesis is respired in this zone (Aristegui *et al.* 2005).

The bathypelagic zone shows several differences compared with upper ecosystems. Considering physical parameters, this system is more stable: water are generally well oxygenated (although anoxic basins exist), temperature exhibits a very narrow range globally, from 1 to 4 ºC, and salinity is practically constant at 35 ppm. In the bathypelagic region the pressure is really high (5 to 10 MPa), but this is not limiting the development of life at macro and micro scale. Despite this apparent homogeneity, it is still possible to recognize several different water masses based on the physical and chemical parameters, being the most important the North Atlantic Deep Water (NADW), the Circumpolar Deep Water (CDW) and the Weddel Sea Deep Water (WSDW). These are found principally in the Atlantic, Pacific and Indian Oceans, respectively.

From a chemical point of view, the concentration of organic matter, inorganic nutrients and other chemical compounds can be very different in different marine regions, depending on the sinking material from the surface. Generally, the bathypelagic ocean is rich in the oxidized forms of inorganic nutrients (NO_3, PO_4) and is depleted of reduced compounds such as ammonium (Nagata *et al.* 2010). Globally the deep ocean is considered the largest reservoir of bioavailable organic carbon (Libes 1992; Benner 2002), and the concentration of dissolved organic carbon (DOC) differs among the different basins (Hansell and Carlson 1998). The role of deep ocean as inorganic carbon sink is quite intuitive (Figure 5a). Between 5 and 15% of the carbon fixed by photosynthesis in the upper marine layer sinks to the bathypelagic realm through the biological pump (Giering *et al.* 2014), where is respired and sequestered for centuries until returned to the upper ocean and then to the atmosphere. Thus, the bathypelagic system has an extremely important role in the global balance of CO_2 and, considering the link of this balance with critical problems such as global warming and climate change, is really important to define the final destination of deep DOC (Aristegui *et al.* 2009).

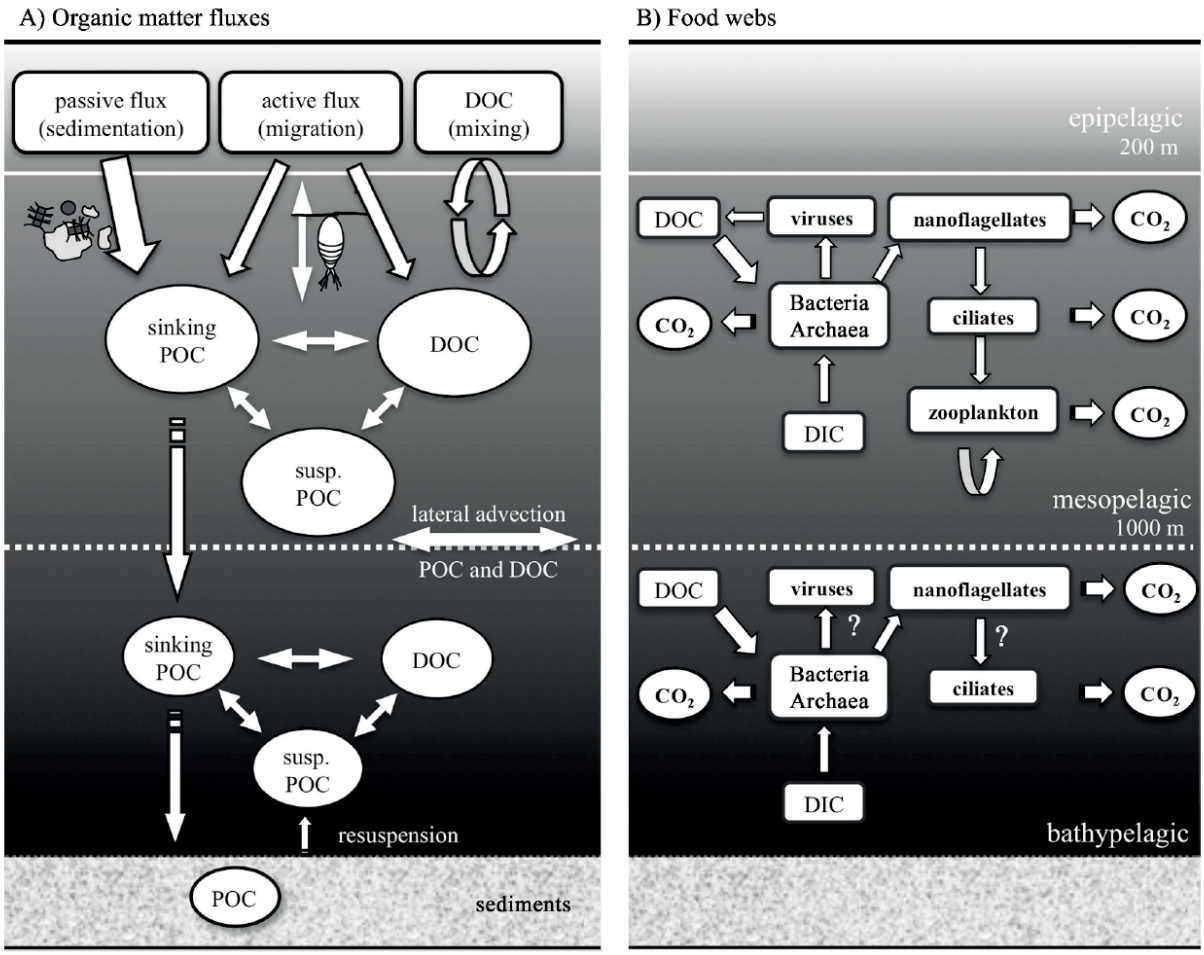


Figure 5. Schematic representation of organic matter fluxes (A) and the microbial food web (B) in the oceanic deep ecosystem (from Aristegui *et al.* 2009). A) Three interconnected pools of organic carbon are indicated: dissolved organic carbon (DOC), sinking particulate organic carbon (POC) and suspended POC. B) Microbial trophic web of the mesopelagic and bathypelagic realms. Prokaryotes in the dark ocean may take up DOC (heterotrophy) and inorganic carbon (chemosynthesis). In the bathypelagic zone prokaryotic control by flagellates or viruses and the role of ciliates remain enigmatic (question marks).

At the bathypelagic level there is a general decrease of DOC concentration along the path of the deep global thermohaline circulation (Figure 6). The concentration of DOC is high (>50 µmol kg^{-1} C) in the newly formed North Atlantic Deep Water (north of 50ºN), tends to be a bit lower and constant in equatorial regions (about 45 µmol kg^{-1} C), and further decreases in the south to a minimum of 39 µmol kg^{-1} C. The constant DOC concentration along South Indian Ocean (40 µmol kg^{-1} C) suggests a net carbon input, due to the invasion of circumpolar deep water (CDW), and then a subsequent removal. Bottom waters of the Pacific Ocean gradually lose organic carbon as they move northward: DOC is 42 µmol kg^{-1} C in the circumpolar waters of south Pacific and decreases to 36 µmol kg^{-1} C as the water slowly enters the deep North Pacific (Figure 7, Hansell *et al.* 2009). Probably this decrease of the DOC is explained by biological consumption by heterotrophic prokaryotes and perhaps fungi.

The abiotic characteristics of the deep ocean defines a habitat really different from the surface one.

There are fragmentary information about the abundance and distribution of microeukaryotes in the dark ocean water column (Tanaka and Rassoulzadegan 2002, Yamaguchi *et al.* 2004, Fukuda *et al.* 2007, Sohrin *et al.* 2010, Morgan-Smith *et al.* 2011, Morgan-Smith *et al.* 2013). Diversity studies have been performed in the water column (López-Garcia *et al.* 2001, Stoeck *et al.* 2003, Countway *et al.* 2007, Not *et al.* 2007b), in sediments (Edgcomb *et al.* 2011, Salani *et al.* 2012) and in marine chimneys of hydrothermal vents (Edgcomb *et al.* 2002, Sauvadet *et al.* 2010). From these papers we know that the diversity of deep water protists appears dominated by Alveolata and

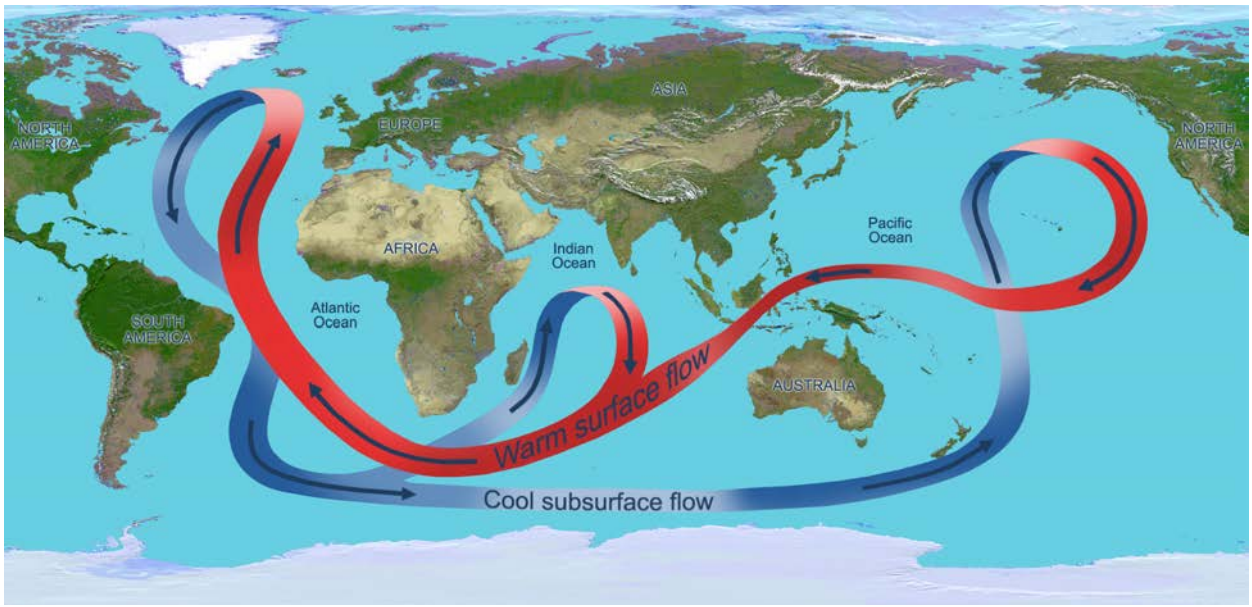


Figure 6. Thermohaline circulation. Cold and dense water masses sink in the North Atlantic and Southern oceans, creating a current that flows in the ocean basins. These waters return to the surface thanks to upwelling events in Indian and North Pacific oceans, forming a current of warm water that flows in the opposite direction in upper layers. Near the North Pole the water get colder and sink restarting a cycle that lasts centuries.

Radiolaria whereas Fungi dominate in sediments. Despite not being one of the dominant groups, Excavata apparently prefer deep waters than surface. As phototrophy is not possible in the dark ocean, the community of bathypelagic protists should present one the three previously mentioned heterotrophic lifestyles: phagotrophy, osmotrophy or parasitism. The relative importance of each trophic mode at the ecosystem perspective is still a matter of debate.

To achieve a global vision of the functioning of the deep ocean, the Malaspina circumnavigation cruise was performed in 2010 on board the R/V BIO_Hesperides. This cruise started in Cadiz (Spain) and sampled 147 stations all over the world (Figure 8). The principal aim of this expedition was the study of the dark ocean at a global scale, including data about microeukaryotes. The magnitude and multidisciplinarity of the sampling effort allowed us to compare our data with parallel parameters in order to achieve a more complete understading of the entire system.

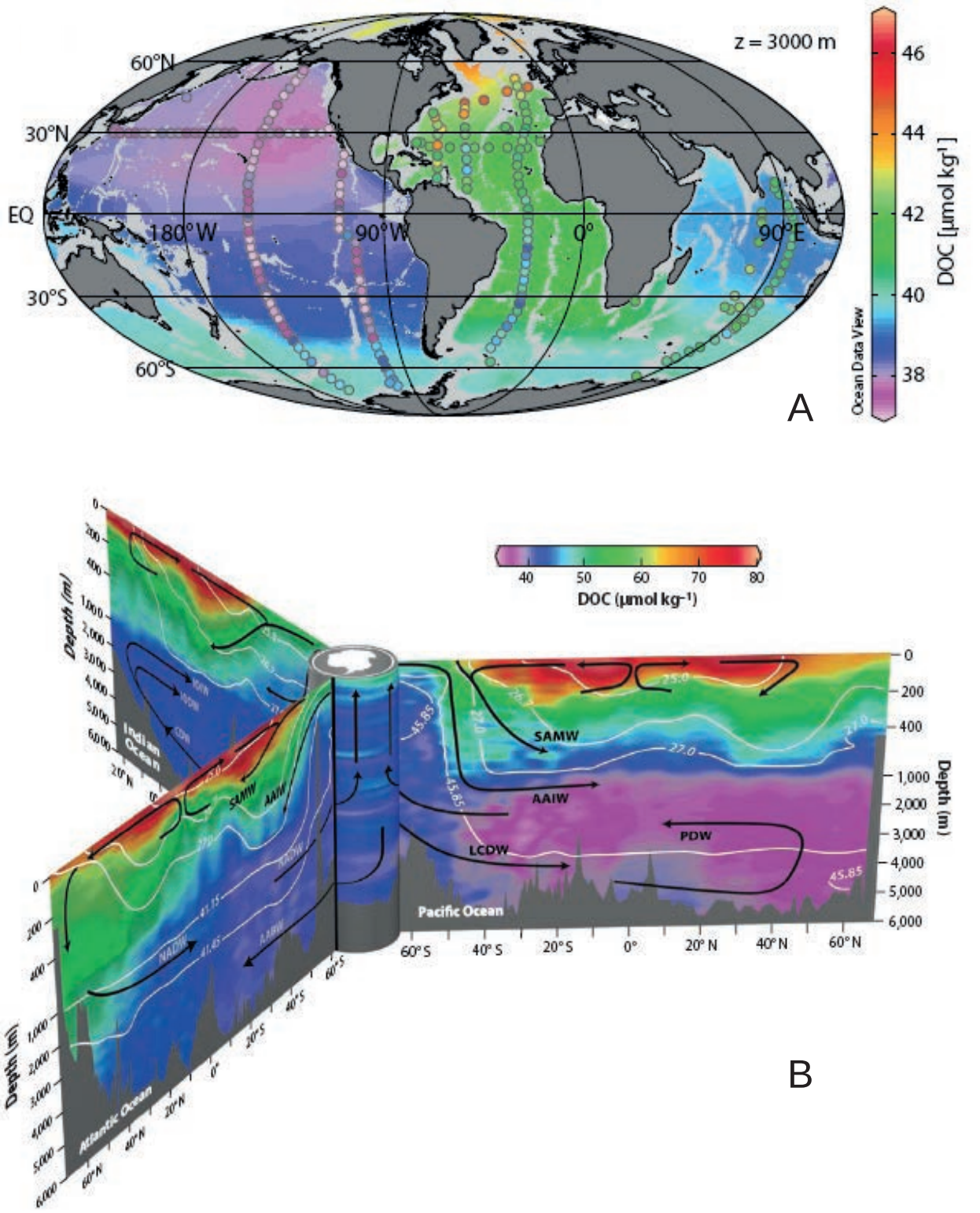


Figure 7. Distribution of dissolved organic carbon (DOC; μmol Kg^{-1}) in the global ocean (Hansell *et al.* 2009). A) Distribution of DOC at 3000 m. Dots are observed values, while the background field is modelled. B) Distribution of DOC in the central Atlantic, central Pacific and eastern Indian ocean. Arrows depict water mass circulation.

Aim of the thesis

The general aim of this thesis is to draw a global picture of the community of marine microeukaryotes. The achievement of this goal was structured in four chapters. The first chapter (*Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean*, ISME J. 2011), the study of the diversity of epipelagic community through clone libraries, was useful as a first approach to molecular biology tools and to establish a guideline on how to treat sequence datasets (i.e. alignment, clustering threshold, diversity estimates). In the second chapter (*General patterns of diversity in major marine microeukaryote lineages*, PLOSONE 2013) sequences derived from all the reports published before 2010 were analyzed in order to describe several features of the genetic diversity of microeukaryotes groups. Moreover, an explorative study of

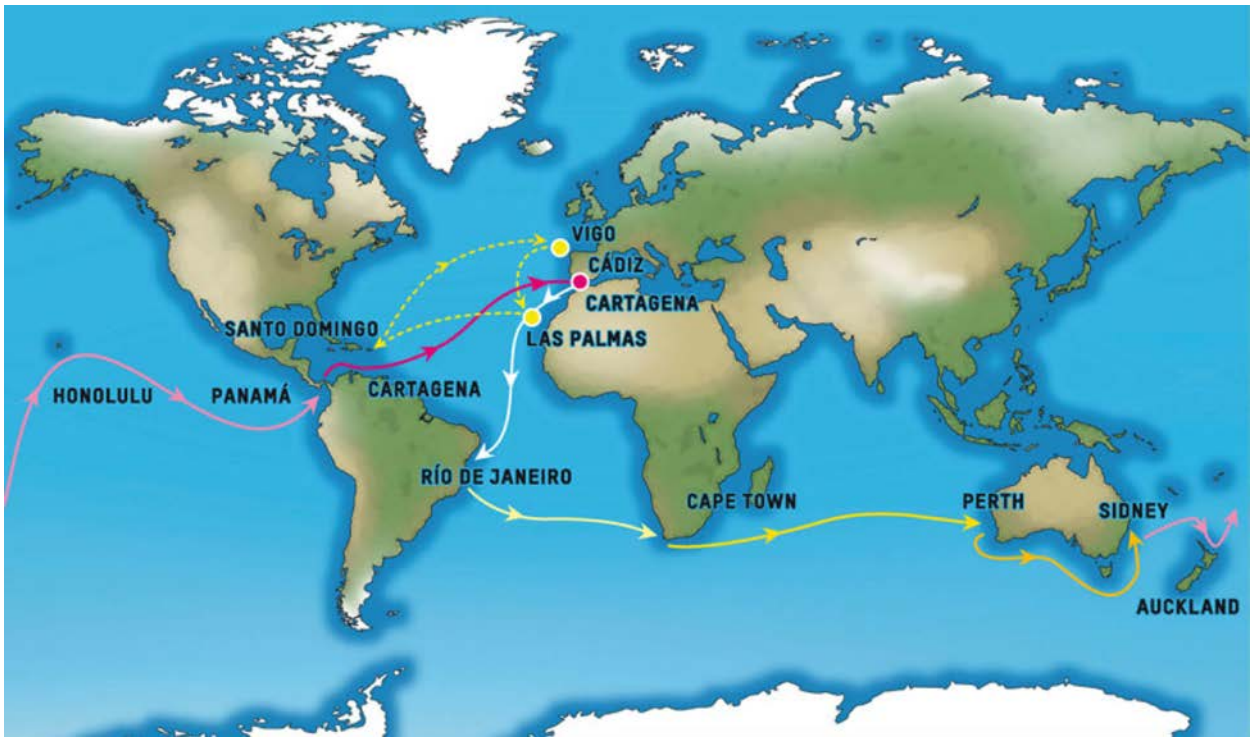


Figure 8. Cruise itinerary of the Malaspina 2010 expedition, including the tracks from the ship Hesperidés (continuous line) and Sarmento de Gamboa (dotted line).

the evolutionary model for the different taxa was performed. The most precious fruit of this work was a well annotated set of sequences, all belonging to the V4 region of 18S rDNA, which were the core for a reference database (MAS9013) used for taxonomic identification and chimera check in the successive studies done by pyrosequencing the same rDNA region. The second part of the thesis, in the frame of Malaspina-2010 project, was focused on the deep ocean ecosystem. The third chapter (*Global abundance of planktonic heterotrophic protists in the deep ocean*, submitted to ISME J.) investigates the abundance of heterotrophic flagellates in the global meso- and bathy-

pelagic regions with the combined use of epifluorescence microscopy and flow-cytometry. In the fourth chapter (*Diversity of marine microeukaryotes in the global deep ocean*, in preparation), we studied the phylogenetic diversity and biogeography of microeukaryotes, and their relation with environmental parameters at the boundary between bathypelagic and abyssal regions through rDNA pyrosequencing and metagenomic approaches.

The outline of different topics studied can be explained under two general objectives and several specific ones, as follows:

Objective 1: *Defining the taxonomic groups of marine microeukaryotes and their genetic structure*

The first part of the thesis represents an effort to define our "diversity unit" from studies based on the well-known molecular cloning and Sanger sequencing in order to establish a solid base for the second part of the thesis. We started with data from one cruise (Chapter 1) and then continued with the analysis of the complete 18S rDNA database available at that time (Chapter 2). The specific objectives of this part were:

- To select the region of the 18S rDNA gene that best represents the variability of the complete gene

- To identify a reasonable similarity threshold for OTU clustering

- To establish the maximal distance in groups delimited at a class-rank level

- To highlight the typical taxonomic classes forming surface communities

Objective 2: *A descriptive study of global deep ocean communities*

The Malaspina expedition allowed us to have a comprehensive set of samples coming from all over the world with associated abiotic and biotic parameters. Such a large amount of data was the base for studying deep microeukaryotes (Chapters 3 and 4) following the next specific objectives:

- To determine the abundance, biomass and distribution of microeukaryotes in the water column between 200 and 4000 m depth

- To study the diversity of bathypelagic microeukaryotes through pyrosequencing and metagenomics approaches

- To identify the abiotic and biotic parameters explaining the abundance and diversity of deep microeukaryotes, with a particular emphasis on the relation with prokaryotes

Chapter 1

Sequences diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean

Massana R, **Pernice M**, Bunge JA, del Campo J (2011). Sequences diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean. *ISME J* **3:** 588-596

Abstract

Despite the ecological importance of marine pico-size eukaryotes, the study of their *in situ* diversity using molecular tools started just a few years ago. These studies have revealed that marine picoeukaryotes are very diverse and include many novel taxa. However, the amount and structure of their phylogenetic diversity and the extent of their sequence novelty still remains poorly known, since a systematic analysis has been seldom attempted. Here we use a coherent and carefully curated dataset of 500 published 18S rDNA sequences to quantify the diversity and novelty patterns of picoeukaryotes in the Indian Ocean. Our phylogenetic tree showed many distant lineages. We grouped sequences in OTUs (Operational Taxonomic Units) at discrete values delineated by pair-wise Jukes-Cantor (JC) distances and tree patristic distances. At 0.01 distance, the number of OTUs observed (237/242; using JC or patristic distances, respectively) was half the number of sequences analyzed, indicating the existence of microdiverse clusters of highly related sequences. At this distance level, we estimated 600-800 OTUs using several statistical methods. The number of OTUs observed was still substantial at higher distances (39/82 at 0.20 distance) suggesting a large diversity at high-taxonomic ranks. Most sequences were related to marine clones from other sites and many were distant to cultured organisms, highlighting the huge culturing gap within protists. The novelty analysis indicated the putative presence of pseudogenes and of truly novel high-rank phylogenetic lineages. The identified diversity and novelty patterns among marine picoeukaryotes are of great importance for understanding and interpreting their ecology and evolution.

Introduction

Planktonic protists play fundamental roles in the functioning of marine ecosystems, both as primary producers and microbial grazers (Sherr *et al*., 2007). Early marine biologists were amazed by the large protist diversity in the plankton, a habitat apparently homogeneous and with a limited range of resources. This phenomenon was named the paradox of the plankton (Hutchinson, 1961). Today it is assumed that biological and environmental factors interact continually, so the plankton habitat never reaches an equilibrium, preventing competitive exclusion by a single species and promoting diversity (Scheffer *et al*., 2003). Little was known for the smallest protists (picoeukaryotes, cells of 0.8-3 μm), which are hardly visible by inverted microscopy. Epifluorescence and flow cytometry counts (Johnson and Sieburth, 1982; Olson *et al*., 1985) revealed their abundance, ubiquity, and ecological relevance, but still did not allow identification. This was made possible with the introduction of molecular tools to oceanography that provided a culturing and microscopic independent assessment of microbial diversity (Giovannoni *et al*., 1990). A series of seminal studies showed that marine picoeukaryotes were indeed very diverse, similar to what was observed for larger protists, and contained many novel lineages (Díez *et al*., 2001; Moon-van der Staay *et al*., 2001; López-García *et al*., 2001). Comparable patterns were also observed in the first molecular surveys of freshwater systems (Lefranc *et al*., 2005; Richards *et al*., 2005).

The methodological improvements to retrieve phylogenetically informative genes from the environment have been paralleled by a growing understanding of the eukaryotic tree of life based on cultured organisms. Phylogenetic analyses have confirmed the taxonomic groups based on cell ultrastructure studies. In addition, phylogenomic analyses have identified a few supergroups composed by eukaryotes with little morphological resemblance but a common evolutionary origin (Baldauf, 2003). The eukaryotic tree of life was first delineated with eight supergroups, which have been further reduced to six (Simpson and Roger, 2004), or less (Burki *et al*., 2008). For instance, the supergroup stramenopiles includes lineages as disparate as the diatoms, chrysophytes or bicosoecids, and the supergroup opisthokonts includes the choanoflagellates, fungi and metazoans. The eukaryotic tree of life represents an optimal framework to assign environmental sequences to known lineages or to define new ones if environmental sequences do not find a place. Thus, novel groups such as marine stramenopiles (MAST, Massana *et al*., 2004), marine alveolates (MALV, Guillou *et al*., 2008), or picobiliphytes (Not *et al*., 2007) have been defined based on environmental surveys. It has been demonstrated that some members of these previously unnoticed lineages are ubiquitous marine grazers, parasites and algae, respectively.

Despite the numerous molecular surveys of marine picoeukaryotes (reviewed in Massana and Pedrós-Alió, 2008; Vaulot *et al*., 2008), the knowledge about the extent of their diversity at different phylogenetic scales and the pattern of sequence novelty (i.e. how different are the environmental sequences from a given study with respect to GenBank sequences) is still in its infancy. Few studies have reported the number of lineages observed grouping sequences at different clustering levels (Caron *et al*., 2009). Parametric and non-parametric statistics have been used to estimate the total richness in different habitats, including picoeukaryotes from the marine plankton (Brown *et al*., 2009). Moreover, little has been advanced in quantifying and representing the novelty patterns of sequences from environmental surveys. Here we are addressing these issues by using a coherent and curated dataset of environmental sequences of picoeukaryotes (500 sequences of ~800 bp). These sequences were just assigned to broad taxonomic groups in a general publication on small protists from the Indian Ocean (Not *et al*., 2008), so the diversity and novelty analyses proposed here are totally new. Specific questions are: How many described taxonomic groups are detected? How many OTUs (Operational Taxonomic Units) are observed when clustering sequences at different thresholds? Is the clustering method affecting the previous question? How many OTUs can be estimated? What is the novelty pattern of environmental sequences? Our study is an effort to describe the diversity and novelty of marine picoeukaryotes exploiting the data gathered in a classical 18S rDNA clone library approach, in order to set up a baseline in which to compare the massive amount of data that are just beginning to be available by high-throughput sequencing (Amaral-Zettler *et al*., 2009; Brown *et al*., 2009; Stoeck *et al*., 2009).

Material and Methods

Sequence dataset

Sequences derive from a recent study in the Indian Ocean (Not *et al*., 2008). Eight clone libraries of the 18S rDNA genes from the picoplankton (0.2 to 3 µm) were prepared from surface and Deep Chlorophyll Maximum samples from stations 01, 09, 18 and 23 (see Fig. 1 in Not *et al*., 2008). Station 01 was coastal, whereas the other three stations (representing 91% of the sequences) were offshore. Details of DNA extraction, PCR (with eukaryotic primers EukA and EukB) and cloning protocols can be found in the original publication. Clones were sequenced with the internal primer 528f, resulting in 572 sequences of around 850 bp. The taxonomic affiliation of each sequence (including chimera detection) was done by BLAST (Altschul *et al*., 1997) and KeyDNATools (http://www.keydnatools.com/) searches and comparison with published phylogenetic trees. A

final dataset of 500 protist sequences was obtained after excluding 30 metazoan sequences, 33 chimeras, and 9 sequences shorter than 500 bp or of low quality. All chromatograms were visually inspected to minimize sequencing errors.

Phylogenetic analysis

Sequences were aligned with MAFFT using the slow and iterative refinement method FFT-NS-i (Katoh *et al*., 2002). The alignment was checked manually and edited with Seaview 3.2 (Galtier *et al*., 1996) to keep the longest region common in most sequences. The final alignment had 961 positions and ~815 bp per sequence (the average size was 797 bp, indicating that most positions in the alignment were covered). Maximum likelihood (ML) phylogenetic trees were constructed with RAxML (Stamatakis, 2006) using the evolutionary model GTR+G+I that best fits our data following ModelTest (Posada and Crandall, 1998). Phylogenetic analyses were done in the freely available University of Oslo Bioportal (www.bioportal.uio.no). Repeated runs on distinct starting trees were carried out to select the tree with the best topology (the one having the best Likelihood of 1000 alternative trees). Bootstrap ML analysis was done with 1000 pseudo-replicates. Trees were edited with the online tool iTOL (Letunic and Bork, 2007).

Grouping sequences in OTUs

Pair-wise Jukes-Cantor (JC) distances among all sequences were computed with PAUP (Swofford, 2002) using an alignment with unique sequences (398 sequences). The distance matrix was processed with DOTUR (Schloss and Handelsman, 2005) to group sequences in OTUs (Operational Taxonomic Units) at different clustering distances. We used the rule of furthest neighbor and the highest precision (p=10000). Heatmaps and Venn diagrams to compare samples were done with the related application Mothur (http://www.mothur.org/). OTUs were also delineated with the online tool RAMI (Pommier *et al*., 2009), which grouped sequences based on their patristic distances (branch lengths). Rarefaction analyses were performed on both DOTUR and RAMI applications using the alignment with all 500 sequences.

Estimating the total number of OTUs

The total number of OTUs (defined at discrete clustering levels) was estimated applying a set of statistical models to the observed OTU abundance. Parametric methods apply a model to the frequency distribution of OTUs and then project the distribution to estimate how many OTUs have been missed (Jeon *et al*., 2006), whereas non-parametric methods such as Chao1 or ACE just apply a simple equation (Chao and Lee, 1992). Several parametric and non-parametric estimators (under different competing models and assumptions) were run at every possible right-truncation

point of the frequency-count data, i.e., omitting outliers (highly abundant taxa in the sample) with the beta version of the program CatchAll built at the Department of Statistical Science, Cornell University. The best parametric model was selected as the one providing the best compromise with a high goodness of fit, low standard error, and maximal use of high frequency counts. The non-parametric method was chosen based on the coefficient of variation of the estimate (Shen *et al.*, 2003).

Novelty analysis

Two values were recorded based on a BLAST search of each environmental sequence against the nucleotide collection (nr/nt) database of NCBI (search on March 2010). The first value was the similarity with the closest environmental sequence in the BLAST output list (Similarity CEM [Closest Environmental Match]), excluding clones from the same library or study. The second was the similarity with the closest cultured organism (Similarity CCM [Closest Cultured Match]), which was the first entry in the list that was taxonomically classified. In a few cases, environmental sequences were so divergent that BLAST calculated the similarity using only a fragment, overestimating the similarity value. This occurred in 21 cases with the CEM and 38 cases with the CCM. In these instances, environmental and GenBank sequences were aligned with MAFFT, and the similarity was calculated using the uncorrected p-distance computed in PAUP. The novelty analysis reported the similarities of the environmental sequences against CEM and CCM in histograms or in dispersion plots (del Campo and Massana, submitted).

Results

Phylogenetic reconstruction of the diversity of marine picoeukaryotes

A maximum likelihood phylogenetic tree with all 500 sequences provided a detailed picture of the diversity of Indian Ocean picoeukaryotes (Figure 1). In this tree, colored branches and external rings are based on the classification of sequences using BLAST and KeyDNATools. Both independent approaches, tree phylogeny and BLAST/ KeyDNATools classification, were remarkably concordant (Figure 1a). The main supergroups (inner ring) were well represented and were divided into taxonomic groups roughly at the Class level (outer ring), most of them with high bootstrap values. Alveolates (dark gray in the inner ring) accounted for most clones in the dataset, in particular dinoflagellates, MALV-I and MALV-II (47% of clones). Stramenopiles (light gray in the inner ring) followed in clonal abundance (19% of clones) and were dominated by

several MAST lineages, chrysophytes and bicosoecids. Rhizaria (blue in the inner ring) were formed mostly by radiolarians (13% of clones). Two cercozoan sequences were closer to ciliates than radiolarians, representing the only example of obvious incorrect phylogenetic placement. Archaeplastida (red in the inner ring) were formed exclusively by prasinophytes and accounted for 4% of clones. The remaining groups (white in the inner ring) contained few badly resolved sequences, including typical marine groups such as haptophytes, cryptophytes, katablepharids, picobiliphytes, telonemida and choanoflagellates. The same tree with real branch lengths (Figure 1b) gave a general impression of the unequal variability contained in each taxonomic group.

This highly supported tree was pivotal to place very divergent sequences that could not be identified by BLAST and KeyDNATools searches (21 sequences shown in light green branches). These novel sequences showed very long branches in the tree (Figure 1b) and, interestingly, some affiliated within a given taxonomic group (marked with an asterisk in Figure 1b). Thus, two divergent sequences were related to MALV-II, one to cercozoans, two to picobiliphytes and one to MAST (the three first cases supported by high bootstrap values). Nevertheless, fifteen novel sequences could still not be related to any taxonomic group not even to a supergroup and occupy highly unique branches in this phylogenetic analysis.

Number of OTUs observed at varying clustering distances

Identical sequences were removed resulting in 398 unique sequences that represented the number of OTUs at null distance. Unique sequences were then grouped into OTUs at distinct thresholds based on JC pair-wise distances and patristic distances displayed in the ML tree. The number of OTUs showed the largest decrease with the initial clustering relaxation (Figure 2a). Thus, the initial 398 OTUs were reduced to 237/242 (JC/Patristic grouping) at 0.01 distance (equivalent to 99% similarity; Figure 2a), meaning that 40% of the unique sequences collapse at this low distance (Figure 2b). We tested that this phenomenon occurred in all phylogenetic groups. After this dramatic initial decline, the number of OTUs continuously decrease when increasing the clustering distance. JC and patristic distances grouped OTUs similarly up to a distance of 0.10, and above this value patristic distances delineated more OTUs (Figure 2a). This cannot be caused by the evolution model, since pair-wise JC and ML distances gave similar values (slope=1.0253; R^2=0.9993; 500 sequences). Instead, these differences appear when the distances are calculated based on the phylogenetic tree. For instance, at a distance of 0.20 roughly separating taxonomic Classes, JC distances delineate 39 OTUs whereas patristic distances delineated 82. These differences are also evident in the distribution of OTUs in distance classes (Figure 2b).

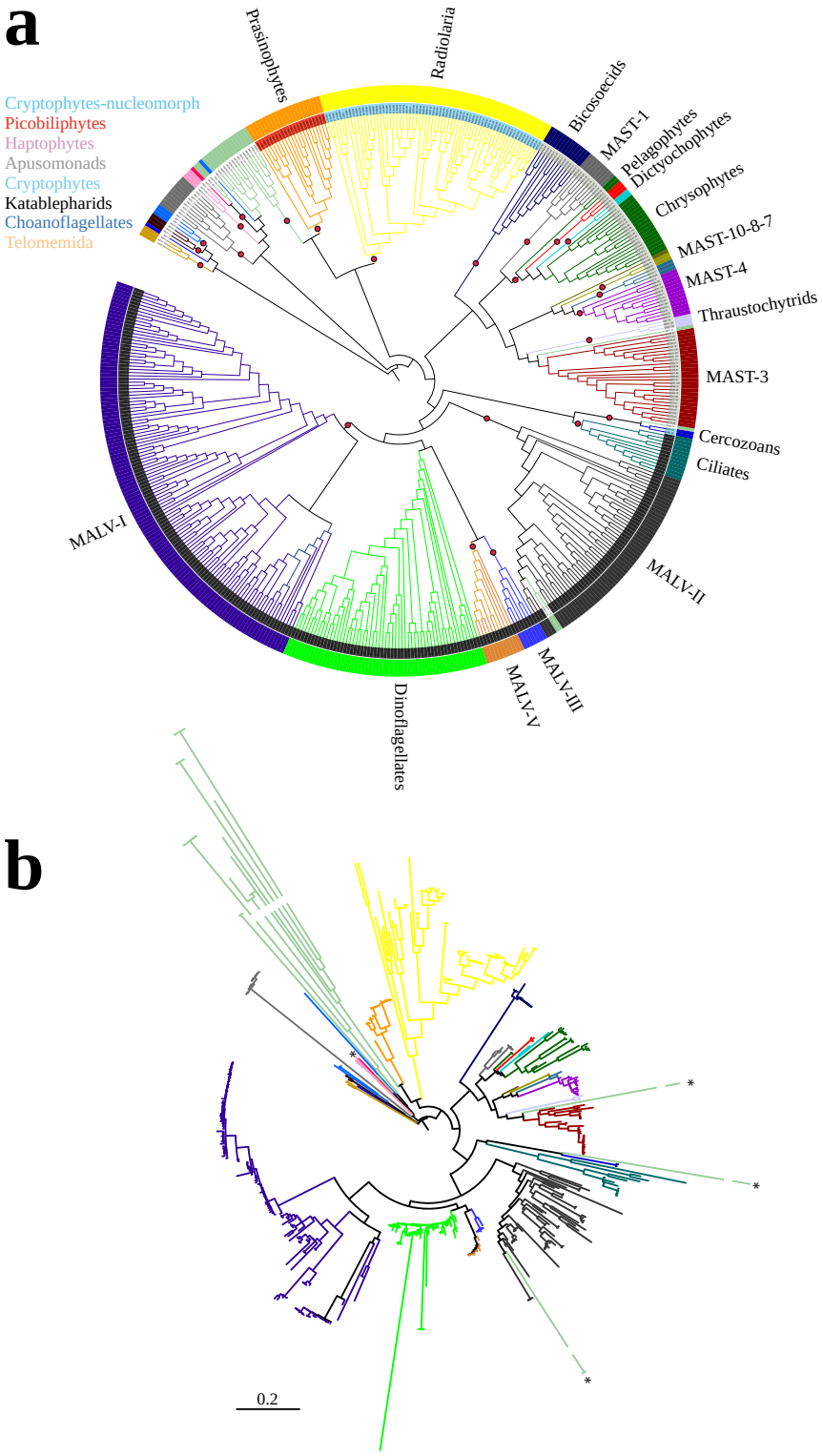


Figure 2 (**a**) Number of OTUs observed after grouping the 398 unique sequences from the Indian Ocean at different clustering levels based on Jukes-Cantor or patristic distances. The correspondence between JC distance and sequence similarity is shown at the top of the graph for comparative purposes. (**b**) Distribution of the number of OTUs in distance classes for both grouping approaches. The area in each class represents the difference in OTUs observed at the two limits of the class (so the OTUs decrease when relaxing the clustering conditions between the two limits). (**c**) Rarefaction curves (OTUs observed versus clones analyzed) at discrete clustering distance levels (from 0.00 to 0.30) for both grouping approaches.

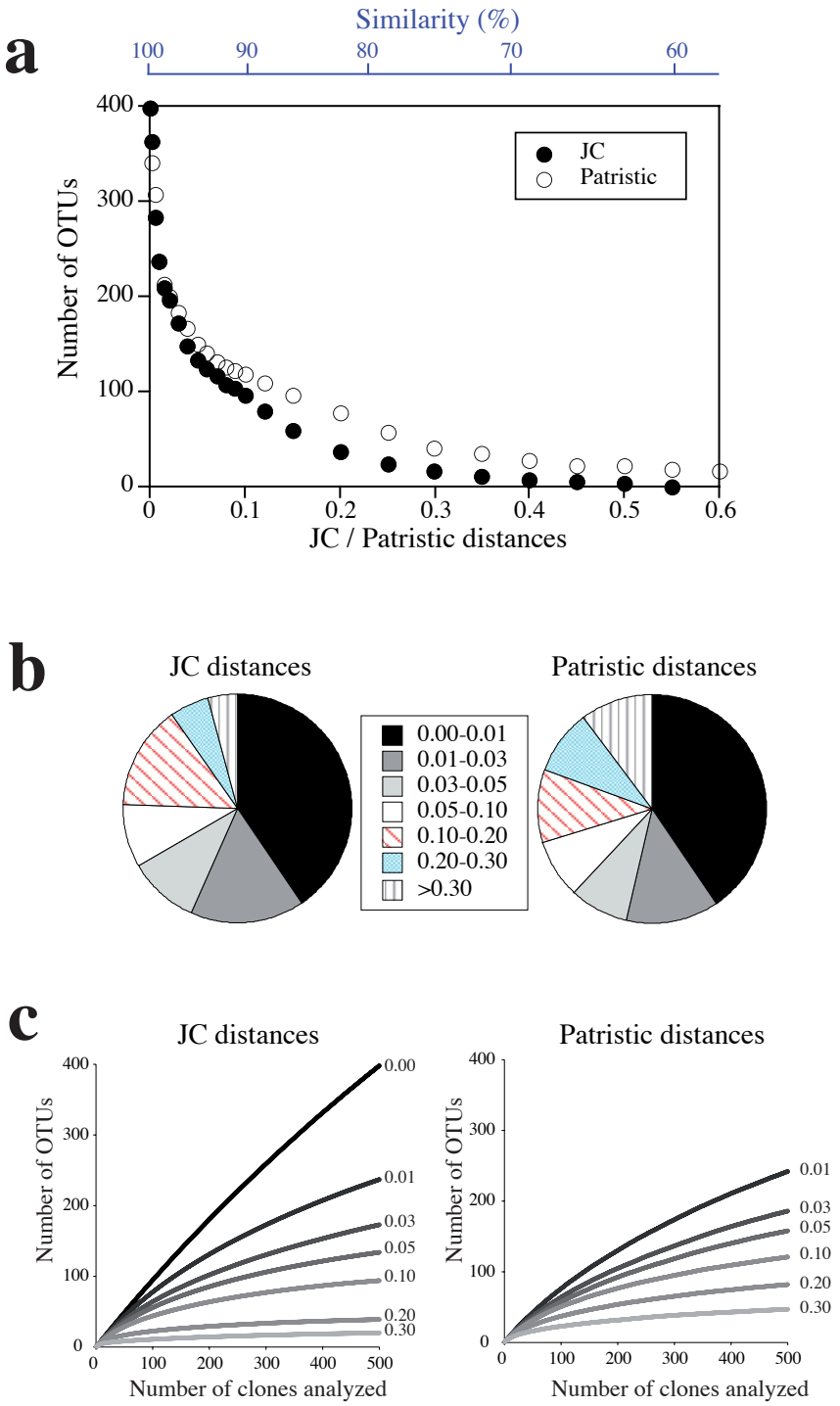


Figure 2 (**a**) Number of OTUs observed after grouping the 398 unique sequences from the Indian Ocean at different clustering levels based on Jukes-Cantor or patristic distances. The correspondence between JC distance and sequence similarity is shown at the top of the graph for comparative purposes. (**b**) Distribution of the number of OTUs in distance classes for both grouping approaches. The area in each class represents the difference in OTUs observed at the two limits of the class (so the OTUs decrease when relaxing the clustering conditions between the two limits). (**c**) Rarefaction curves (OTUs observed versus clones analyzed) at discrete clustering distance levels (from 0.00 to 0.30) for both grouping approaches.

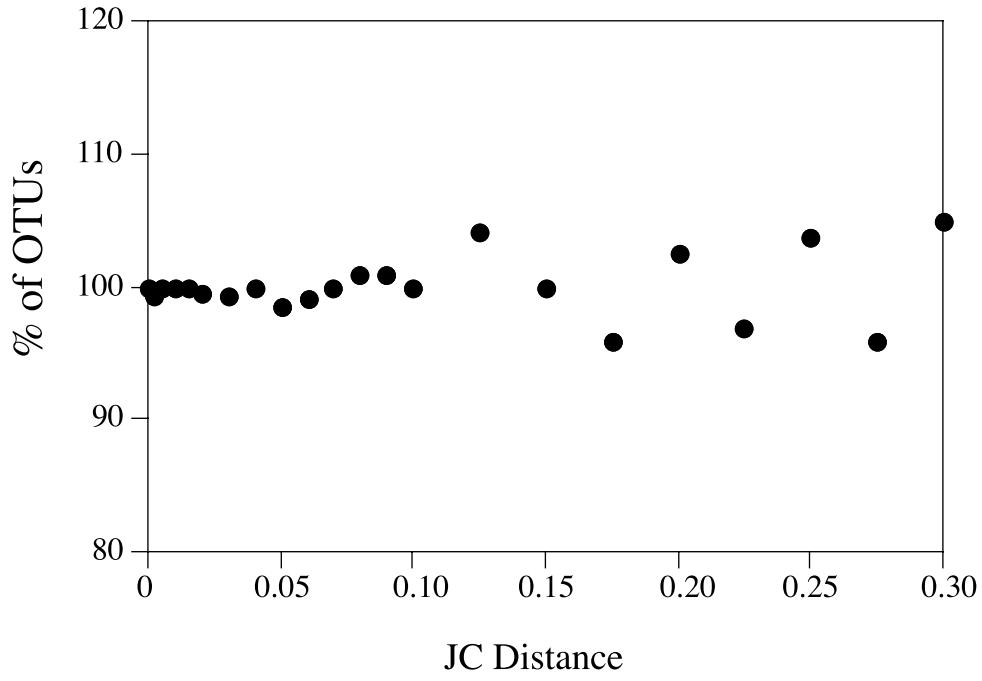


Figure 3 Percentage of total OTUs (estimated with the whole dataset) that are recovered in 23 analyses with defined phylogenetic groups and adding up the counts for each separate group. This comparison was done at 22 discrete clustering distance levels.

Rarefaction curves were then constructed to relate the number of OTUs to the sequencing effort. The rarefaction curve with OTUs grouped at null JC distance did not show any sign of saturation (Figure 2c). Rarefaction curves done with OTUs clustered at increasing distances showed a progressively better coverage, with plateaus starting to be evident at levels of 0.2 and 0.3 for both grouping methods. This indicated a severe undersampling to retrieve OTUs defined stringently, but at the same time suggested that the higher-rank phylogenetic groups were moderately well represented in the sequence dataset.

Our dataset included a huge sequence variability (Figure 1b) and raised doubts about the accuracy of the alignment used for calculating pair-wise distances and the ML tree. In addition, hypervariable regions, kept to report the variability at all scales, were inevitably ambiguously aligned. Thus, we expected that doing separate analyses for coherent phylogenetic groups would yield better OTU counts. We prepared sequence datasets with the taxonomic groups shown in Figure 1a and redid the OTU counting (alignment, JC distances and DOTUR) for the 23 separate sets. Then, the number of OTUs in each set were added up and compared with the number observed with the whole dataset. To our surprise, both approaches gave similar OTU numbers at clustering distance levels up to 0.30, being almost identical at all levels tested up to 0.10 (Figure 3). This exercise sustains the accuracy of the OTU counts and the ML tree obtained using the whole, and very variable, dataset.

Table 1 Observed and estimated number of OTUs defined at discrete clustering levels (based on JC and patristic distances) within the 500 sequences of picoeukaryotes from the Indian Ocean

	JC - distance grouping							Patristic distance grouping						
Distance	Observed	Parametric estimate			Nonparametric estimate			Observed	Parametric estimate			Nonparametric estimate		
0.00	**398**	**1951**	*193*	SE	**1320**	*162*	AC							
0.01	**237**	**731**	*150*	ME	**609**	*91*	A1	**242**	**803**	*188*	ME	**700**	*117*	A1
0.02	**197**	**624**	*160*	ME	**552**	*96*	A1	**205**	**617**	*120*	ME	**646**	*122*	A1
0.03	**173**	**472**	*116*	ME	**396**	*64*	A1	**186**	**710**	*311*	ME	**685**	*155*	A1
0.04	**149**	**312**	*45*	ME	**243**	*25*	AC	**170**	**557**	*175*	ME	**440**	*79*	A1
0.05	**134**	**251**	*34*	ME	**203**	*19*	AC	**158**	**486**	*151*	ME	**394**	*73*	A1
0.07	**117**	**224**	*33*	ME	**176**	*18*	AC	**135**	**357**	*84*	ME	**306**	*56*	A1
0.10	**94**	**158**	*21*	ME	**129**	*12*	AC	**121**	**257**	*49*	ME	**223**	*35*	A1
0.12	**78**	**147**	*26*	ME	**132**	*24*	A1	**113**	**231**	*37*	ME	**177**	*20*	AC
0.15	**58**	**91**	*15*	ME	**77**	*9*	A1	**99**	**184**	*22*	ME	**151**	*18*	AC
0.20	**39**	**61**	*14*	ME	**61**	*15*	A1	**82**	**159**	*25*	ME	**151**	*29*	A1
0.30	**20**	**27**	*4*	SE	**35**	*14*	A1	**47**	**85**	*16*	ME	**93**	*26*	A1

The estimated number of OTUs was calculated under several parametric and nonparametric methods, showing the estimated value (bold), the standard error (italics) and the best fitting model or index (SE: Single Exponential; ME: Two Mixed Exponential; AC: ACE; A1: ACE1).

Number of OTUs estimated at varying clustering distances

The rarefaction curves clearly showed that our dataset underestimated diversity, particularly whenOTUs were defined at low genetic distances. In order to estimate the "total" number of OTUs, we applied several statistical methods on their frequency distribution (Table 1). Parametric models tend to predict higher estimates than nonparametric indices, and this was also observed here. The best parametric estimate obtained at null distance was 1951 (± 193), and a distance of 0.01 was 731 (±150; JC-grouping) or 803 (±188; Patristic grouping). OTU estimates at increasing distances decrease parallel to the decrease in the observed number, although observed and estimated values get closer at high distances. Whereas at 0.01 the observed OTUs represent 32-30% of the estimated value, at 0.20 they represent 63-51% (Table 1). So we are missing many more low-rank taxa than high-rank lineages.

Novelty analysis of marine picoeukaryotes

For each sequence, the similarity against the closest environmental match (CEM) and the closest cultured match (CCM) was recorded. The average CEM similarity (97.9%) was much higher than the average CCM similarity (91.9%). The similarity distribution against CEM was skewed towards the highest values, with a marked peak at 99%, whereas CCM similarity distributed well from 85% to 100%, with minor peaks at 87%, 92% and 99% (Figure 4a). A dispersion plot of both similarity

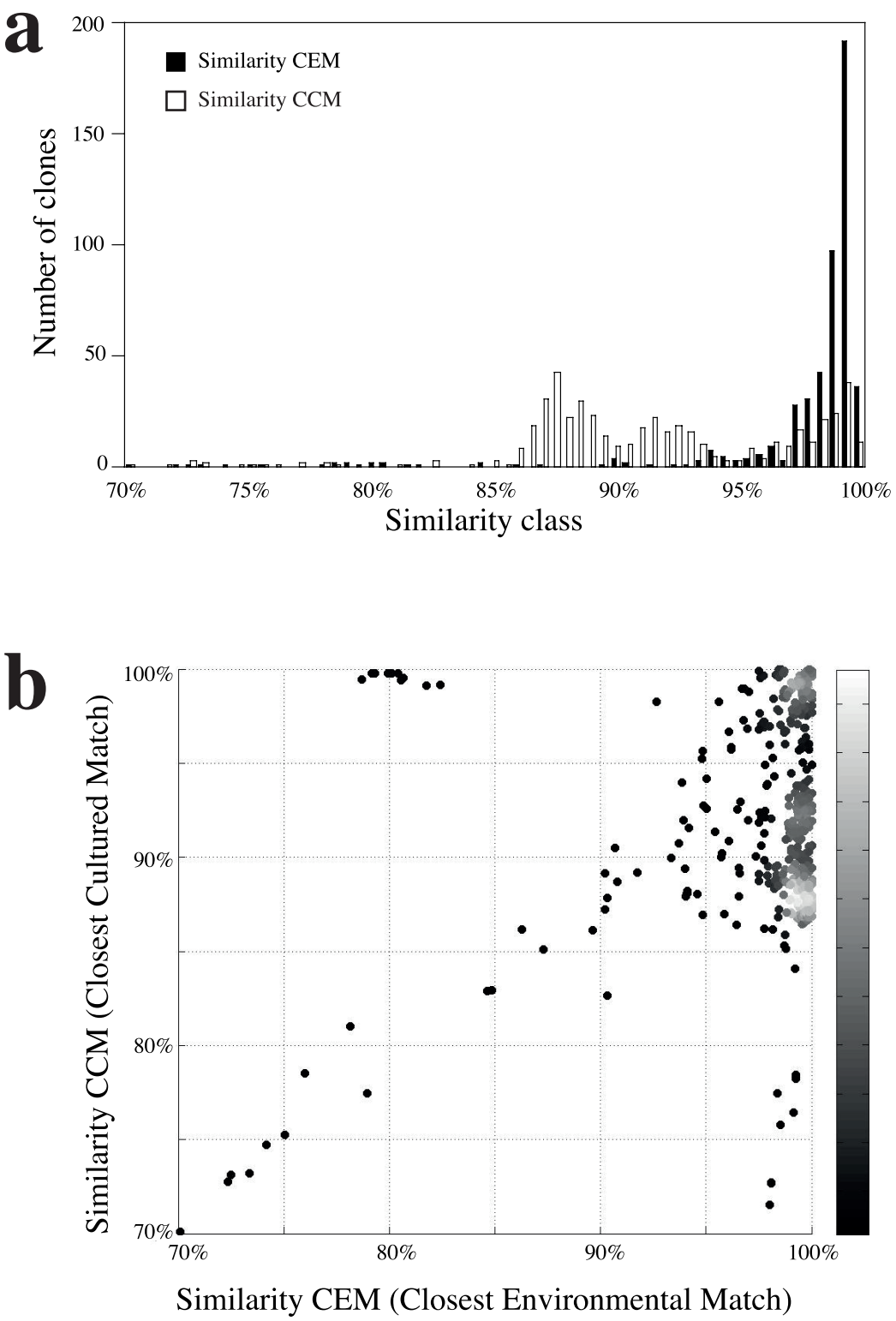


Figure 4 Novelty analysis of the 500 sequences of picoeukaryotes retrieved from the Indian Ocean. (**a**) Histogram showing the distribution of similarities against Closest Environmental Match (CEM) and Closest Cultured Match (CCM) of all sequences, in 0.5% similarity classes. (**b**) Dispersion plot of the CEM and CCM similarities for each sequence, with dots shaded depending on the number of neighbors (light gray dots indicate a dense area, whereas black dots a disperse area).

values showed that few sequences were closer to CCM than to CEM, with most dots at the 1:1 line or below (Figure 4b). A notable exception were ten sequences close to *Amastigomonas debruynei* but only 80% similar to a marine clone. Dots were shaded depending the neighbors they have, unveiling two dense areas (Figure 4b). The first was limited by CEM and CCM similarities above 98% (17% of sequences) and included sequences close to cultured organisms and marine clones. The second dense area was limited by CEM scores above 98% and CCM similarities between 87 and 93% (42% of sequences) and included sequences close to marine clones but distant to cultured organisms. The plot also highlighted novel sequences. Dots below 80% similarity in both axis indicated very divergent sequences never found before. Some sequences were only 75% similar with all sequences in GenBank except a few marine clones. Thus, three related clones were 98% similar to a single sequence from the Mediterranean Sea, whereas three other clones were 95-99% similar to a few Sargasso and Mediterranean Seas sequences.

Each particular phylogenetic group might exhibit a different novelty pattern, as exemplified with the supergroups alveolates and stramenopiles (Figure 5). In both cases most sequences placed in the area with high CEM similarities (>98%) and had a particular behavior with respect to CCM. Thus, some stramenopile groups are at the top of the graph with high CCM scores (bicosoecids, dictyochophytes, pelagophytes), chrysophytes show an intermediate position with 90-95% CCM similarities, whereas MASTs and thraustochytrids have CCM similarities below 90% (Figure 5a). A similar distribution can be described for alveolates, with dinoflagellates at the top of the graph, followed by MALV-III and -V at an intermediate position, and MALV-I and -II with lowest CCM scores (Figure 5b).

Comparing the diversity among samples

The protist composition in different samples was compared using their OTU content defined at 0.01 distance, roughly corresponding to species, and at 0.20 distance, roughly corresponding to a taxonomic level of Class. Data was displayed in heatmaps that quantify the pair-wise difference among samples, and Venn diagrams that show the number of unique and shared OTUs. At low distance, samples strongly differed among each other (Figure 6a), as expected due to the undersampling shown in rarefaction curves and statistical estimates. Still, some ecologically sound information was derived from the maps: the coastal sample was the most different, and the closest pairs were two offshore surface samples (58 and 70) and two offshore DCM samples (33 and 72). Venn diagrams were then done to compare coastal, surface and DCM samples. At low distance level, only a few OTUs were shared and unique OTUs were as high as 64% (coastal), 78% (surface) and 72% (DCM). As expected, OTUs grouped at a higher distance gave a different

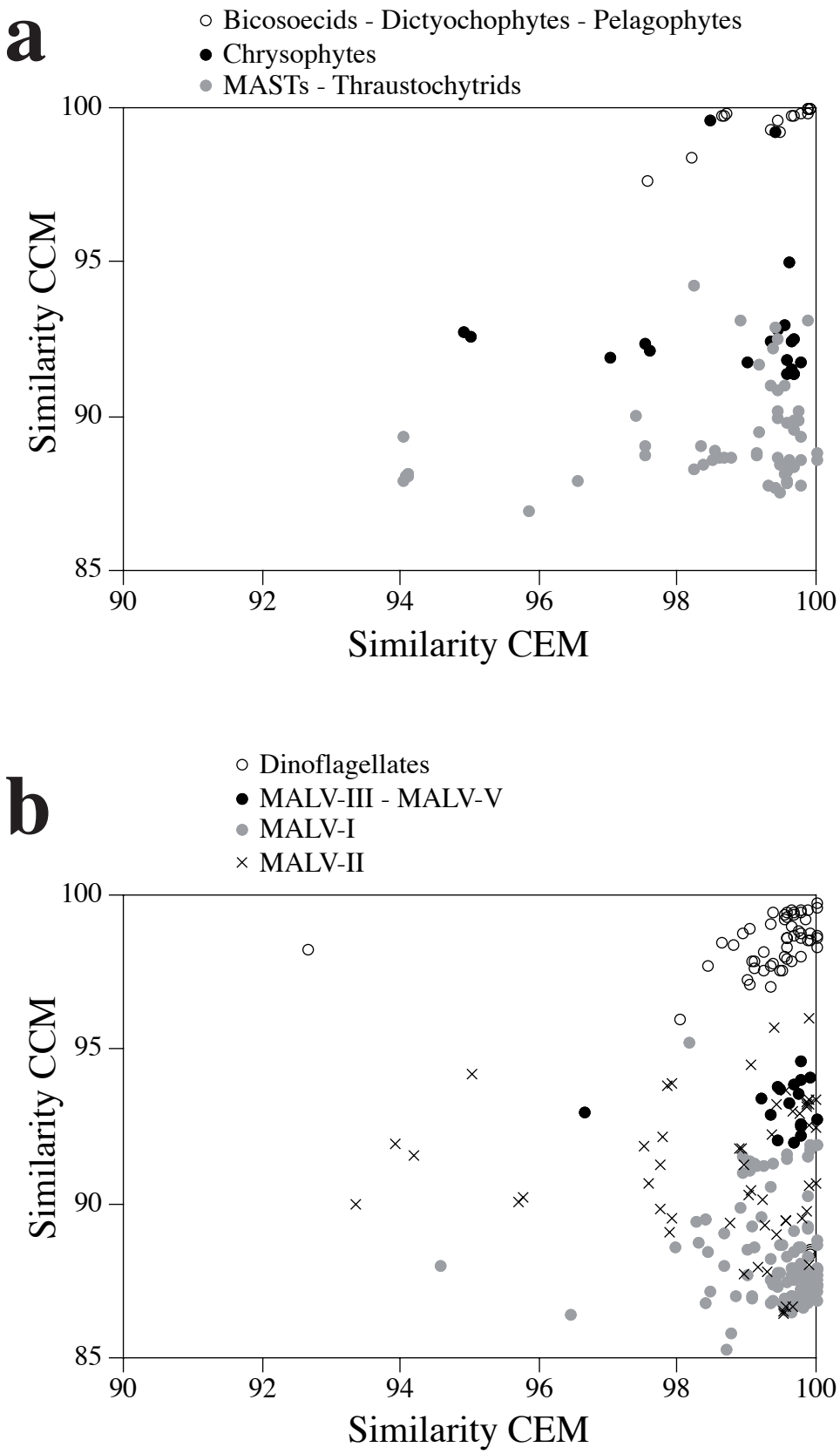


Figure 5 Dispersion plot of the CEM (Closest Environmental Match) and CCM (Closest Cultured Match) similarities for sequences affiliating to stramenopiles (**a**) and alveolates (**b**) separated in several taxonomic groups.

picture (Figure 6b). Heatmaps showed a homogenization of samples and Venn diagrams showed fewer unique OTUs (8% in coastal; 37% in surface; 33% in DCM), suggesting a rather coherent high-rank diversity among the samples analyzed.

Discussion

Taxonomic groups detected

We used a dataset of five-hundred 18S rDNA sequences published before (Not *et al*., 2008) to describe and quantify the diversity and novelty of picoeukaryotes from the Indian Ocean. 18S rDNA sequences were not complete (they were almost half of the gene, >800 bp), so they could contain insufficient positions for sound phylogenies. Moreover, it was not clear whether an alignment including very divergent sequences could retrieve the proper relationships among them. The alignment was first used for a ML phylogenetic tree, which recovered the main supergroups and most taxonomic groups (Figure 1). In fact, the tree-independent sequence classification (via BLAST and KeyDNATools) was concordant with the tree. Second, we compared the OTU number computed from the whole alignment or by adding the values of 23 separate alignments. Again, the results were satisfactory, as minor differences were found at all clustering levels tested (Figure 3). These exercises indicated that MAFFT could deal with very variable sequence inputs and that our partial sequences were long enough for proper phylogenies, as was shown for bacterial 16S rDNA partial sequences (Stackebrandt and Rainey, 1995). These tests add consistency to the results presented here.

The ML tree displayed a large diversity at different phylogenetic scales, pointing out that the seemingly homogeneous picoeukaryotic assemblages seen by epifluorescence microscopy or flow cytometry are formed by cells with very divergent evolutionary histories. As in all studies based on size-fractionated biomass, it is possible that some of these sequences do not derive from picoeukaryotes but from larger cells broken during the filtration or detrital DNA, so the true picoeukaryote diversity we present here might be overestimated. The high-rank diversity observed here, both in terms of eukaryotic supergroups detected and the presence and relative abundance of specific lineages, was typical of molecular surveys of marine picoeukaryotes (Massana and Pedrós-Alió, 2008; Vaulot *et al*., 2008). Alveolates and stramenopiles were the most common groups. Rhizaria and archaeplastida appeared on a second level and were represented by a single lineage each. A unique choanoflagellate sequence (fungi were absent) represented the opisthokonts,

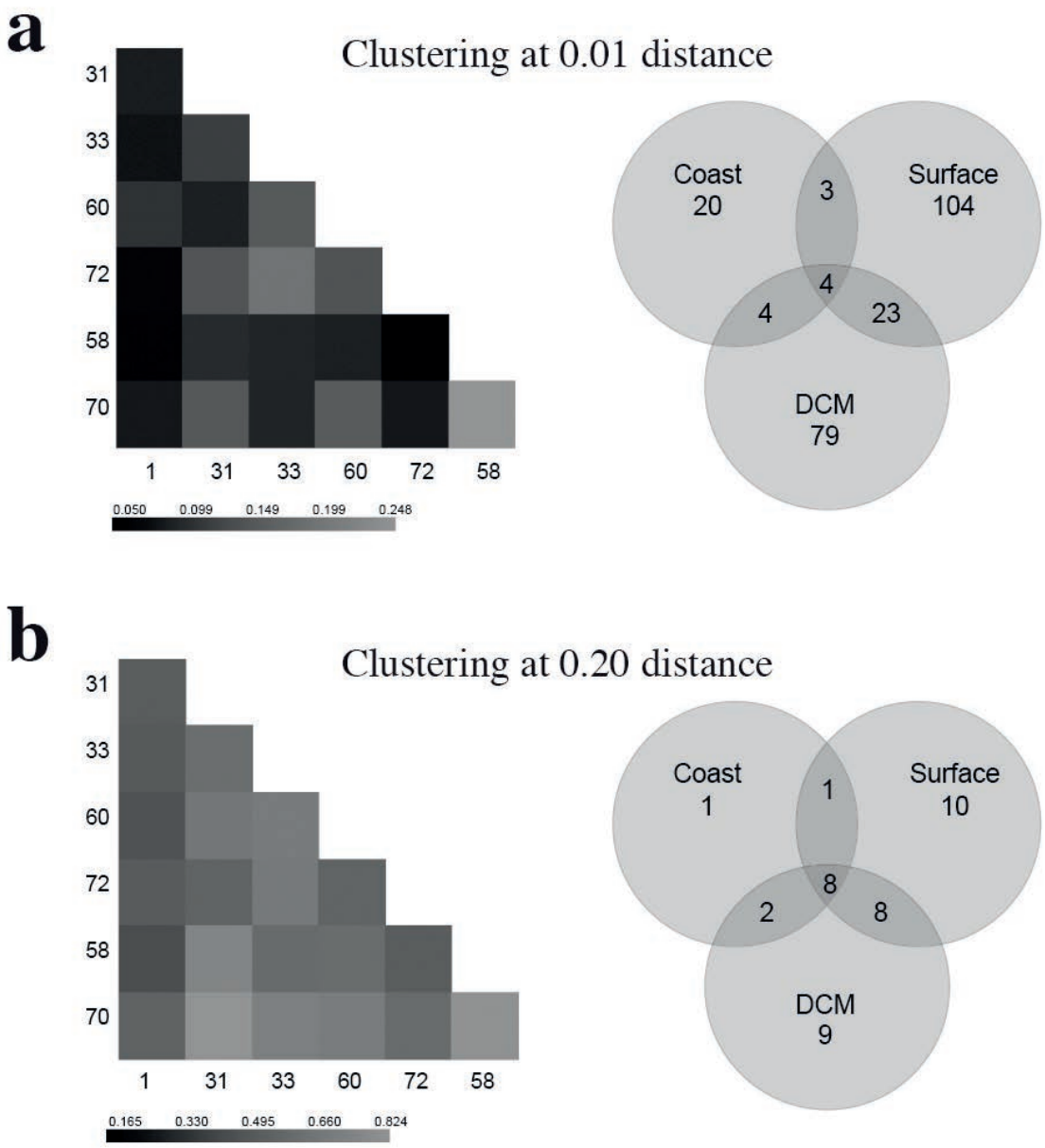


Figure 6 Heatmaps (left) and Venn diagrams (right) comparing the diversity of marine picoeukaryotes among samples, using the shared or unique OTUs defined at clustering JC distances of 0.01 (**a**) or 0.20 (**b**). Samples derive from the coast (1), offshore surface (31, 58 and 70) or offshore DCM (33, 60 and 72).

whereas excavates and amoebozoa were not detected in this particular dataset. Several reasons might explain the absence (or very low abundance) of some lineages in our libraries. First, some groups, such as many excavates, are likely unable to thrive in the marine plankton. Second, some cells could be excluded during the prefiltration step, such as larger loricated choanoflagellates (Leakey *et al.*, 2002) or particle-living amoebas (Rogerson *et al.*, 2003). Third, some lineages could be too scarce to be detected (Pedrós-Alió, 2007), a concern that can be partially solved by high-throughput sequencing, or by enriching the sample with the cells of interest using flow

cytometry cell sorting (Shi *et al*., 2009). Finally, some lineages could remain undetected due to inefficient DNA extraction or biased PCR amplification (Wintzingerode *et al*., 1997). This will only be solved by modifying DNA extraction protocols and applying new universal and group-specific PCR primers.

Observed and estimated richness at different clustering levels

An important issue when quantifying the diversity of a natural assemblage is how to define the countable units. Ideally, units are biological species, which work reasonably well for macroorganisms but are impractical in the microbial world, particularly within picoplankton, where diversity is determined using DNA sequence data. In order to create tractable units, sequences above a given distance threshold are pragmatically grouped into Operational Taxonomic Units (OTUs). When done at discrete clustering levels, this provides the number of OTUs at different phylogenetic scales and yields information on the genetic structure of microbial assemblages (Acinas *et al*., 2004; Shaw *et al*., 2008). This analysis has seldom been done with marine picoeukaryotes (Caron *et al*., 2009). The most stringent criteria using null distance would be supported by laboratory studies that show that only strains with identical rDNA gene sequences are sexually compatible (Amato *et al*., 2007). In our dataset, only 20% of sequences are not contributing to a new OTU at null distance, highlighting the large diversity of the dataset.

The largest decrease in OTU number occurred with the initial relaxation of the clustering conditions. This OTU collapse was caused by the presence of a substantial number of very similar (≥ 99%) but seldom identical sequences. This microdiversity could be explained by a combination of methodological, biological and ecological factors. First, PCR or sequencing errors might account for part of these minute differences. In our dataset, chromatograms were visually inspected to confirm the high quality of the reads and remove ambiguous positions, so few sequencing errors would be expected. Second, the rDNA gene in eukaryotes appears typically in tandem repeats varying from a few to several thousand copies depending the taxa (Zhu *et al*., 2005). Copies are generally homogenized by concerted evolution (Dover, 1982), but this process is not always complete and minor differences can be found within the same genome (Alverson and Kolnick, 2005). Third, in absence or low frequency of sexual reproduction, a plausible scenario for many protist species (Weisse, 2008), marine picoeukaryotes could experience similar evolutionary processes as bacteria and reveal equivalent microdiverse clusters (Acinas *et al*., 2004). These clusters would be generated by neutral mutations (their genetic and functional diversity would be neutral), and could be regarded as natural taxonomic units or ecological species (Cohan, 2006).

The number of OTUs kept decreasing when increasing distances. At distances up to 0.10, the grouping using JC or patristic distances showed a good correspondence, whereas above 0.10 both clustering methods deviate significantly, with patristic distances delineating more OTUs at a given clustering level. This is the expected and described trend (Pommier et al., 2009), and occurs because patristic distances among two sequences, especially if they are divergent, are systematically larger than JC distances. OTUs produced by patristic distances are based on genetic change and would result in a more accurate and evolutionary robust clustering, but this is not yet a common practice in microbial ecology. The high number of OTUs detected at large distances results from the combination of a remarkable high-rank diversity (many taxonomic groups and supergroups) and the presence of very long branches at different positions of the tree (within well defined groups or forming novel high-rank lineages).

Comparing observed and estimated OTU values allowed evaluating the undersampling of our data. Parametric estimators, which are known to work better with low coverage datasets such as ours (Epstein and López-García, 2008), predicted 1951 OTUs at null distance and 731/803 OTUs at 0.01 distance (JC/Patristic grouping, respectively). Thus, we only retrieved a glimpse of picoeukaryotic diversity (20% and 32-30%, respectively). Increasing the clustering distance level the diversity coverage also increased (consistent with the rarefaction analysis) meaning that we started to miss less lineages. Our estimates ranked among the highest detected in surveys of microbial eukaryotes using clone libraries. At a similarity clustering level of 99% (distance of 0.01), our estimate was higher than the 398 OTUs from marine anoxic samples (Jeon *et al*., 2006), 107 OTUs from hypersaline deep samples (Alexander *et al*., 2009), or 605 OTUs from hydrothermal vent samples (Stoeck *et al*., 2007). In surface marine samples, 572 OTUs were estimated at a clustering level of 95% (Countway *et al*., 2007), a number slightly higher to ours. The unique high-throughput sequencing study with estimates from marine surface samples gave much higher values: 56292 OTUs at 100% similarity, 9231 at 99% and 3765 at 95% (Brown *et al*., 2009). This study was based on early 454 technology, which sequenced a very short amplicon (>50 bp) and could overestimate diversity due to low-frequency errors. Thus, whereas the actual numbers have to be regarded with caution, it seems clear that this study (and the many more to come with improved technologies) will raise significantly the higher limit of protistan diversity. Overall, marine picoeukaryotes appeared as very diverse assemblages.

Novelty analysis of environmental sequences

Novelty of environmental sequences was inferred based on their similarity with the GenBank database. At the time of the first eukaryotic molecular surveys, only the similarity against CCM

could be calculated, yielding generally low values (Díez *et al*., 2001). This situation changed after years of molecular surveys and thousands of deposited sequences. In present studies, marine environmental sequences have generally high CEM scores (with clones from other marine studies) whereas CCM scores still remain low. So, the large sequencing effort on marine picoeukaryotes during the last 10 years has not been paralleled by a significant culturing success, as revealed by the still uncultured MAST or MALV groups. Overall, our data highlight the huge culturing gap existing for the dominant marine picoeukaryotes.

The novelty analysis pointed out very divergent sequences that appeared in the area of the dispersion plot with very low CCM and/or CEM values. These sequences formed very long branches in the ML phylogenetic tree generally with an unresolved position, although some could be robustly placed in a taxonomic group based on the tree (see stars in Figure 1b). Nine sequences showed very low CCM and CEM scores, meaning that they are very distant to any existing sequence. We speculate that these unique sequences could be pseudogenes (Thornhill *et al*., 2007), and this could be confirmed by secondary structure models. If pseudogenes, they would not have any ecological implication and would not stand as separate biological units. Six sequences were extremely divergent from any sequence except to a few marine clones. It is unlikely (but not impossible) that sequences retrieved thousands of kilometers apart are pseudogenes. Instead, these could represent high-rank novel phylogenetic lineages and are obvious candidates for further research. Retrieving additional sequences, constructing sound phylogenies, and visualizing the target cells by FISH will identify if they are truly novel taxonomic units.

Concluding remarks

In this study we explored the diversity and novelty patterns of marine picoeukaryotes using 18S rDNA clone libraries and Sanger sequencing Our observations and the new exploratory approaches presented here can be adapted to facilitate the analysis of the massive amounts of data from the Next Generation Sequencing (NGS) technologies. It should be pointed out that although far from saturation, clone libraries can still provide longer sequences and of very high quality as compared with current NGS methods. We showed here that picoeukaryotes from the Indian Ocean were very diverse at distinct phylogenetic scales. In fact, we are only seeing the tip of the iceberg of their diversity and it is expected that NGS will allow investigating this underexplored space. Our data also indicated microdiverse clusters similar to those found in bacteria, but it is early to explain them by ecological factors or by biological or methodological factors. Most sequences from the Indian Ocean were highly similar to environmental sequences from other marine sites, indicating a widespread distribution of similar lineages, and many were far from cultured organisms, revealing

a significant culturing gap. We also highlighted very divergent sequences, and we speculated that some could be pseudogenes and others could be novel high-rank phylogenetic lineages. From an ecological perspective, our quantitative sequence analysis would help to address fundamental questions of what generates, maintains and structures the large diversity observed, and what are the functional implications of this large diversity at different scales. From an evolutionary perspective, we are faced with very divergent sequences that could account for new, unexpected and fascinating evolutionary lineages.

Acknowledgments

This study was supported by the projects GEMMA (CTM2007-63753-C02-01/MAR, MEC), the NSF grant DEB-0816638 and the European Funding Agencies from the ERA-net program BiodivERsA, under the BioMarKs project. We thank Marco Álvarez for help in BLAST analysis, Miguel Lurgi and Baptiste Mourre for help in MatLab and Jose Castresana and Ramiro Logares for useful advices on phylogeny.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551-554.

Alexander E, Stock A, Breiner H-W, Behnke A, Bunge J, Yakimov MM *et al.* (2009). Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. *Environ Microbiol* **11**: 360-381.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Alverson AJ, Kolnick L. (2005). Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *J. Phycol* **41**: 1248-1257.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009). A method for studying protistan diversity using massively parrallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**: e6372

Amato A, Kooistra WHCF, Ghiron JHL, Mann DG, Pröschold T, Montresor M. (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* **158**: 193-207.

Baldauf SL. (2003). The deep roots of eukaryotes. *Science* **300**: 1703-1706.

Brown MV, Philip GK, Bunge JA, Smith MC, Bisset A, Lauro FM *et al.* (2009). Microbial community structure in the North Pacific ocean. *ISME J* **3**: 1374-1386.

Burki F, Schalchian-Tabrizi K, Pawlowski J. (2008). Phylogenomics reveals a new «megagroup» including most photosynthetic eukaryotes. *Biology Lett* **4**: 366-369.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD *et al.* (2009). Defining DNA-based Operational Taxonomic Units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**: 5797-5808.

Chao A, Lee S-M. (1992). Estimating the number of classes via sample coverage. *J Amer Stat Assoc* **87**: 210-217.

Cohan FM. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Phil Trans R Soc B* **361**: 1985-1996.

Countway PD, Gast RJ, Dennet MR, Savai P, Rose JM, Caron DA. (2007). Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ Microbiol* **9**: 1219-1232.

Díez B, Pedrós-Alió C, Massana R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol* **67**: 2932-2941.

Dover GA. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299:** 111-117.

Epstein S, López-García P. (2008). "Missing" protists: a molecular prospective. *Biodivers Conserv* **17**: 261-276.

Galtier N, Gouy M, Gautier C. (1996). SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**: 543-548.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60-63.

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al*. (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ Microbiol* **10**: 397-408.

Hutchinson GE. (1961). The paradox of the plankton. *Am Nat* **95**: 137-145.

Jeon S-O, Bunge J, Stoeck T, Barger KJA, Hong S-H, Epstein SS. (2006). Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72**: 6578-6583.

Johnson PW, Sieburth JMcN. (1982). In-situ morphology and occurrence of eucaryotic phototrophs of bacterial size in the picoplankton of estuarine and oceanic waters. *J Phycol* **18**: 318-327.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

Leakey RJG, Leadbeater BSC, Mitchell E, McCready SMM, Murray AWA. (2002). The abundance and biomass of choanoflagellates and other nanoflagellates in waters of contrasting temperature to the north-west of South Georgia in the Southern Ocean. *Eur J Protistol* **38**: 333–350.

Lefranc M, Thénot A, Lepère C, Debroas D. (2005). Genetic diversity of small eukaryotes in lakes differing by their trophic status. *Appl Environ Microbiol* **71**: 5935-5942.

Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127-128.

López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603-607.

Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A *et al.* (2004). Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol* **70**: 3528-3534.

Massana R, Pedrós-Alió C. (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213-218.

Moon-van der Staay SY, De Wachter R, Vaulot D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409:** 607-610.

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Töbe K *et al*. (2007). Picobiliphytes: A marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315**: 252-254.

Not F, Latasa M, Scharek R, Viprey M, Karleskind P, Balagué V *et al.* (2008). Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep Sea Res I* **55**: 1456-1473.

Olson RJ, Vaulot D, Chisholm SW. (1985). Marine-phytoplankton distributions measured using shipboard flow-cytometry. *Deep Sea Res* **32**: 1273-1280.

Pedrós-Alió C. (2007). Dipping into the rare biosphere. *Science* **315**: 192-193.

Pommier T, Canbäck B, Lundberg P, Hagström Å, Tunlind A. (2009). RAMI, a tool for identification and characterization of phylogenetic clusters in microbial communitites. *Bioinformatics* **25**: 736-742.

Posada D, Crandall KA. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA. (2005). The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive

and globally dispersed lineages. *Environ Microbiol* **7**: 1413-1425.

Rogerson A, Anderson OR, Vogel C. (2003). Are planktonic naked amoebae predominantly floc associated or free in the water column? *J Plankton Res* **25**: 1359-1365.

Scheffer M, Rinaldi S, Huisman J, Weissing FJ. (2003). Why plankton communities have no equilibrium: solutions to the paradox. *Hydrobiologia* **491**: 9-18.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining Operational Taxonomic Units and estimating especies richness. *Appl Environ Microbiol* **71**: 1501-1506.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JBH. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200-2210.

Shen TJ, Chao A, Lin CF. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**: 798-804.

Sherr BF, Sherr EB, Caron DA, Vaulot D, Worden AZ. (2007). Oceanic protists. *Oceanography* **20**: 130-134.

Shi XL, Marie D, Jardillier L, Scanlan DJ, Vaulot D. (2009). Groups without cultured representatives dominate eukaryotic picophytoplankton in the oligotrophic South East Pacific Ocean. *PLoS ONE* **4**: e7657.

Simpson AGB, Roger AJ. (2004). The real «kingdoms» of eukaryotes. *Curr Biol* **14**: R693-R696.

Stackebrandt E, Rainey FA. (1995). Partial and complete 16S rDNA sequences, their use in generation of 16S rDNA phylogenetic trees and their implications in molecular ecological studies. In: Akkermans ADL, van Elsas JD, de Bruijn FJ (eds). *Molecular microbial ecology manual*. Kluwer Academic: Dordrecht, pp 1-17.

Stamatakis A. (2006). RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* **22**: 2688-2690.

Stoeck T, Kasper J, Bunge J, Leslin C, Ilyin V, Epstein S. (2007). Protistan diversity in the Arctic: A case of paleoclimate shaping modern biodiversity? *PLoS ONE* **2**: e278.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A *et al.* (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology* **7**: 72.

Swofford DL. (2002). PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Mass.

Thornhill DJ, Lajeunesse TC, Santos SR. (2007). Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Mol Ecol* **16**: 5326-5340.

Vaulot D, Eikrem W, Viprey M, Moreau H. (2008). The diversity of small eukaryotic phytoplankton ($\leq 3\ \mu$m) in marine ecosystems. *FEMS Microbiol Rev* **32**: 795-820.

Weisse T. (2008). Distribution and diversity of aquatic protists: and evolutionary and ecological perspective. *Biodivers Conserv* **17**: 243-259.

Wintzingerode Fv, Göbel UB, Stackebrandt E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21**: 213-229.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52:** 79-92.

Chapter 2

General patterns of diversity in major marine microeukaryote lineages

Pernice MC, Logares R, Guillou L, Massana R (2013). General Patterns of diversity in major marine microeukaryote lineages. *PLOS ONE* **8:** e57170, doi: 10.1371/journal.pone.0057170

Abstract

Microeukaryotes have vital roles for the functioning of marine ecosystems, but still some general characteristics of their current diversity and phylogeny remain unclear. Here we investigated both aspects in major oceanic microeukaryote lineages using 18S rDNA (V4-V5 hypervariable regions) sequences from public databases that derive from various marine environmental surveys. A very carefully and manually curated dataset of 8291 Sanger sequences was generated and subsequently split into 65 taxonomic groups (roughly to Class level based on KeyDNATools) prior to downstream analyses. First, we calculated genetic distances and clustered sequences into Operational Taxonomic Units (OTUs) using different distance cut-off levels. We found that most taxonomic groups had a maximum pairwise genetic distance of 0.25. Second, we used phylogenetic trees to study general evolutionary patterns. These trees confirmed our taxonomic classification and served to run Lineage Through Time (LTT) plots. LTT results indicated different cladogenesis dynamics across groups, with some displaying an early diversification and others a more recent one. Overall, our study provides an improved description of the microeukaryote diversity in the oceans in terms of genetic differentiation within groups as well as in the general phylogenetic structure. These results will be important to interpret the large amount of sequence data that is currently generated by High Throughput Sequencing technologies.

Introduction

Decoding the complexity of marine microeukaryotic diversity is one of the biggest challenges of modern microbial ecology, given the astonishingly large diversity detected in molecular surveys [1-6]. Thousands of high-quality environmental Sanger sequences derived from clone libraries of the 18S rDNA genes are now available in public databases, and represent an important resource to investigate some aspects of the general architecture of protist diversity that still remain unclear. Pair-wise distances among environmental sequences are generally used to cluster them into Operational Taxonomic Units (OTUs) at different distance levels. The number of OTUs at each clustering threshold, defined here as "clustering pattern", is a useful proxy of the diversity magnitude and it can also be used to characterize intra group distances. Clustering patterns have already been described for whole protist communities [7-10], but it is expected that the analysis of singular groups can highlight interesting diversity differences among lineages. These features are better reflected in the shape of phylogenetic trees from where we can infer the «phylogenetic structure» of a group, that is, the specific diversification patterns drawn by the branches (number, length and relative positions) of a phylogenetic tree [11]. Very little has been done to investigate these structures in specific groups of marine microbial eukaryotes.

The clustering pattern, based on pair-wise genetic distances, has the advantage of being easily comparable among datasets and strongly related to sequence similarity. Indeed, OTU counts provide an estimate of present diversity in each taxonomic group. Alternatively, the phylogenetic structure derived from the branching pattern of a tree gives a complementary view that contains imprints of evolutionary events occurring within given lineages. The phylogenetic structure is the result of the interplay between speciation and extinction through time, processes that are driven by factors such as geographical isolation, environmental restrictions, reproduction modes and intraspecific interactions [12]. Different protist groups may exhibit different propensities for net rate of cladogenesis (speciation minus extinction rates, [13]) over time [14], and these different evolutionary histories can influence their phylogenetic structure.

An important issue when clustering sequences in OTUs is the meaning of the clustering level applied. Several studies have attempted to identify the threshold fitting species definitions, to establish a countable unit in biodiversity inventories. Sequences sharing a similarity above 98% of the 18S rDNA gene have been proposed to derive from the same species [15,16], but we are far from a general agreement on which value to use. Another fundamental question is identifying the maximum genetic distance that can be contained within a given phylogenetic group, regarded as a collection of species sharing the same evolutionary origin as well as several biological and ecolo-

gical properties. In protist taxonomy, a relevant grouping level is the rank «Class» that targets, for instance, dinoflagellates, diatoms, and choanoflagellates. This analysis will also allow comparing traditional Classes with new ribogroups. The latter emerge from molecular surveys, do not have cultured representatives, and are dispersed throughout the eukaryotic tree of life. Significant ribogroups are the MALV within Alveolata [17], the MAST within Stramenopiles [18], and the RAD within Rhizaria [9].

Here we used publicly available 18S rDNA Sanger sequences obtained from molecular surveys aimed to study the diversity of marine planktonic protists by a culture-independent approach. We classified these sequences into separate taxonomic groups, combining classical taxonomy (Class level) with ribogrouping, and analyzed the genetic diversity in each group by OTU clustering and phylogeny. Our main objective was to get an improved representation of marine protist diversity. This will serve as a frame for interpretation and comparison with data obtained by High Throughput Sequencing (HTS) technologies like 454 or Illumina [19]. HTS sequences (that is, reads) need to be validated against data retrieved independently; otherwise they can produce strongly biased views of diversity [20,21]. In summary, this study allowed us a) to establish the maximum genetic distance value for each taxonomic group, b) to obtain an improved picture of the diversity of different groups, and c) to get an overview of the diversification history within different lineages.

Results

In this study we carried out an analysis of very carefully curated 18S rDNA environmental sequences derived from marine surveys both from oxic and anoxic water samples (see Table S1). A first filtering step retained 13,270 sequences of marine planktonic protists obtained from clone libraries done with universal-eukaryotic primers (Fig. S1). These were classified into 65 taxonomic groups and only sequences containing the V4-V5 regions were kept (8291 sequences; Fig. S2). Some of these groups were well-defined classical taxa (mostly at the class level) whereas the rest were ribogroups deriving exclusively from molecular environmental surveys (Table 1 and Table S2). Alveolata sequences constituted more than half of the dataset, being MALV-II (with 1815 sequences), Dinophyceae, MALV-I and Ciliophora the most represented. Stramenopiles were second in the number of sequences and included more taxonomic groups than Alveolata (21 versus 10). The largest groups within Stramenopiles were Bacillariophyceae, Chrysophyceae, MAST-3 and MAST-1. Rhizaria were represented by 682 sequences, distributed among several cercozoan and radiolarian groups. The recently proposed CCTH supergroup (Cryptophyta,

Table 1. Classification of environmental 18S rDNA sequences in 42 taxonomic major groups.

Supergroup	Group			Distances			OTUs		
			Seq	Avg	Max	Max_c	0.00	0.01	0.05
Opisthokonta	*Choanoflagellatea*	C	100	0.13	0.30	**0.24**	89	**56**	32
Rhizaria	*Acantharea*	C	129	0.15	0.29	**0.26**	110	**63**	29
	Chlorarachniophyceae	C	33	0.14	0.24	**0.23**	29	**13**	7
	Larcopyle	O	18	0.02	0.05	-	13	**4**	1
	Monadofilosa	S	81	0.11	0.30	**0.22**	72	**56**	33
	*Nassellaria**	O	52	0.18	0.41	**0.32**	45	**29**	19
	RAD A	R	37	0.17	0.29	**0.26**	34	**23**	15
	RAD B	R	88	0.11	0.23	**0.16**	66	**36**	17
	Spumellaria	O	209	0.06	0.26	**0.13**	154	**79**	20
Archaeplastida	*Prasinophyceae*	C	551	0.09	0.31	**0.21**	376	**130**	30
	Trebouxiophyceae	C	89	0.01	0.12	**0.04**	26	**11**	6
Stramenopiles	*Bacillariophyceae*	C	253	0.14	0.30	**0.29**	207	**120**	57
	Bicosoecea	C	75	0.11	0.35	**0.28**	60	**34**	17
	Bolidophyceae	C	63	0.05	0.12	**0.11**	34	**12**	7
	Chrysophyceae	C	152	0.13	0.27	**0.24**	115	**75**	32
	Dictyochophyceae	C	91	0.09	0.22	**0.16**	65	**35**	16
	Eustigmatophyceae	C	15	0.01	0.03	-	11	**3**	1
	Labyrinthulida	C	29	0.17	0.35	**0.34**	26	**19**	17
	MAST-1	R	107	0.08	0.20	**0.16**	74	**28**	9
	MAST-2	R	20	0.01	0.05	-	13	**6**	2
	MAST-3	R	149	0.12	0.27	**0.21**	110	**73**	31
	MAST-4	R	92	0.03	0.07	**0.06**	60	**24**	3
	MAST-7	R	82	0.04	0.14	**0.08**	48	**21**	6
	MAST-8	R	17	0.07	0.13	-	14	**9**	6
	MAST-12	R	26	0.16	0.27	-	24	**19**	16
	Oomyceta	C	19	0.11	0.29	-	16	**13**	10
	Pelagophyceae	C	34	0.01	0.07	**0.02**	22	**8**	2
	Pirsonids	-	47	0.03	0.09	**0.08**	37	**26**	5
CCTH	*Cryptophyceae*	C	179	0.09	0.24	**0.21**	130	**45**	3
	Katablepharids	-	20	0.02	0.06	-	12	**6**	2
	Picobiliphyceae	R	53	0.07	0.20	**0.15**	42	**24**	8
	Prymnesiophyceae	C	193	0.08	0.30	**0.14**	148	**90**	37
	Telonemia	C	68	0.05	0.12	**0.11**	60	**42**	9
Alveolata	*Ciliophora*	P	956	0.18	0.42	**0.37**	788	**434**	187
	Dinophyceae	C	1018	0.07	0.50	**0.24**	848	**463**	122
	MALV-I	R	980	0.19	0.48	**0.42**	779	**431**	132
	MALV-II	R	1815	0.16	0.38	**0.30**	1517	**900**	353
	MALV-III	R	79	0.05	0.15	**0.11**	60	**38**	9
	MALV-V	R	51	0.02	0.07	**0.04**	41	**19**	3
Excavata	*Diplonemea*	C	58	0.11	0.21	**0.21**	56	**51**	27
	Kinetoplastea	C	40	0.23	0.39	**0.37**	31	**22**	15
Incertae sedis	*Apusomonadidae*	C	14	0.15	0.41	-	9	**6**	4

Each group is coded according to their taxonomic rank (S: subphylum; C: class; O: order; G: genus; R: ribogroup). The table shows the number of sequences per group (Seq), the average (Avg), maximum (Max) and maximum corrected (Max_c) pair-wise distances, and the number of OTUs at three cut-off levels. *Nassellaria comprises also the order Collodaria

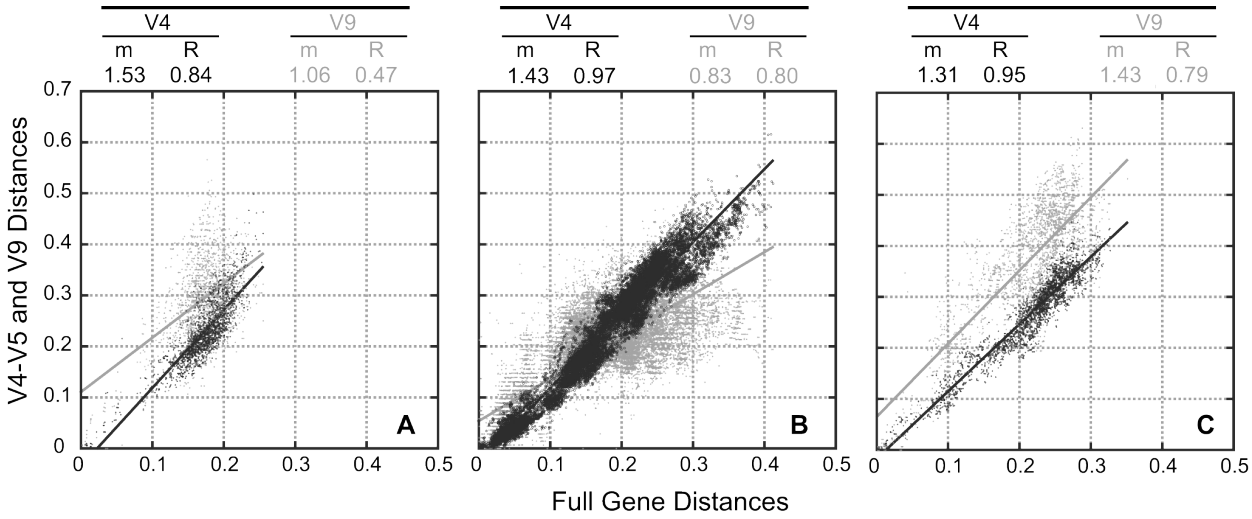


Figure 1. Comparison of partial and full-length 18S rDNA sequences to infer genetic distances. The three panels show pair-wise genetic distances (Jukes Cantor corrected) of the complete gene against partial regions (V4-V5 in dark grey or V9 in light grey) for sequences within Stramenopiles (A), Alveolata (B), and Rhizaria (C). Slopes (m) and coefficients (R) of the correlations are shown at the top of the graphs.

Centroheliozoa, Telonemia, Haptophyta, Burki et al. [22]), was present in the dataset with 522 sequences, mainly from Prymnesiophyceae and Cryptophyceae. The remaining groups contained less than 90 sequences, with the exceptions of Choanoflagellatea and Prasinophyceae. Finally, 427 sequences remained unidentified (could not be assigned to even a supergroup), and were labeled as Novel.

Justifying the target 18S rDNA region

The rationale of choosing the V4-V5 region (~550 bp) for most analyses was to maximize the number of sequences with shared positions, since many clone libraries targeted this region. We investigated how well this partial region represented the variability of the complete 18S rDNA gene. This test also included the V9 region (~160 bp). For the three separate datasets (Stramenopiles, Alveolata and Rhizaria) we plotted the pair-wise distances calculated with the two partial regions (V4-V5 and V9) with respect to the distances computed using the full-length gene (Fig. 1). The V4-V5 region gave better results, with higher correlation coefficients (R) in the three cases (0.84 to 0.97) as compared with the values derived from the V9 region (0.47 to 0.80). In addition, the slopes of the correlation (m) were similar considering the V4-V5 region (1.31 to 1.53) whereas varied largely using the V9 region (from 0.83 to 1.43). So, this indicated that the V4-V5 region (but not the V9 region) represented well the variability of the entire 18S rDNA gene. The V4-V5 region was more variable than the complete gene, overestimating genetic distances by a factor of ~1.4.

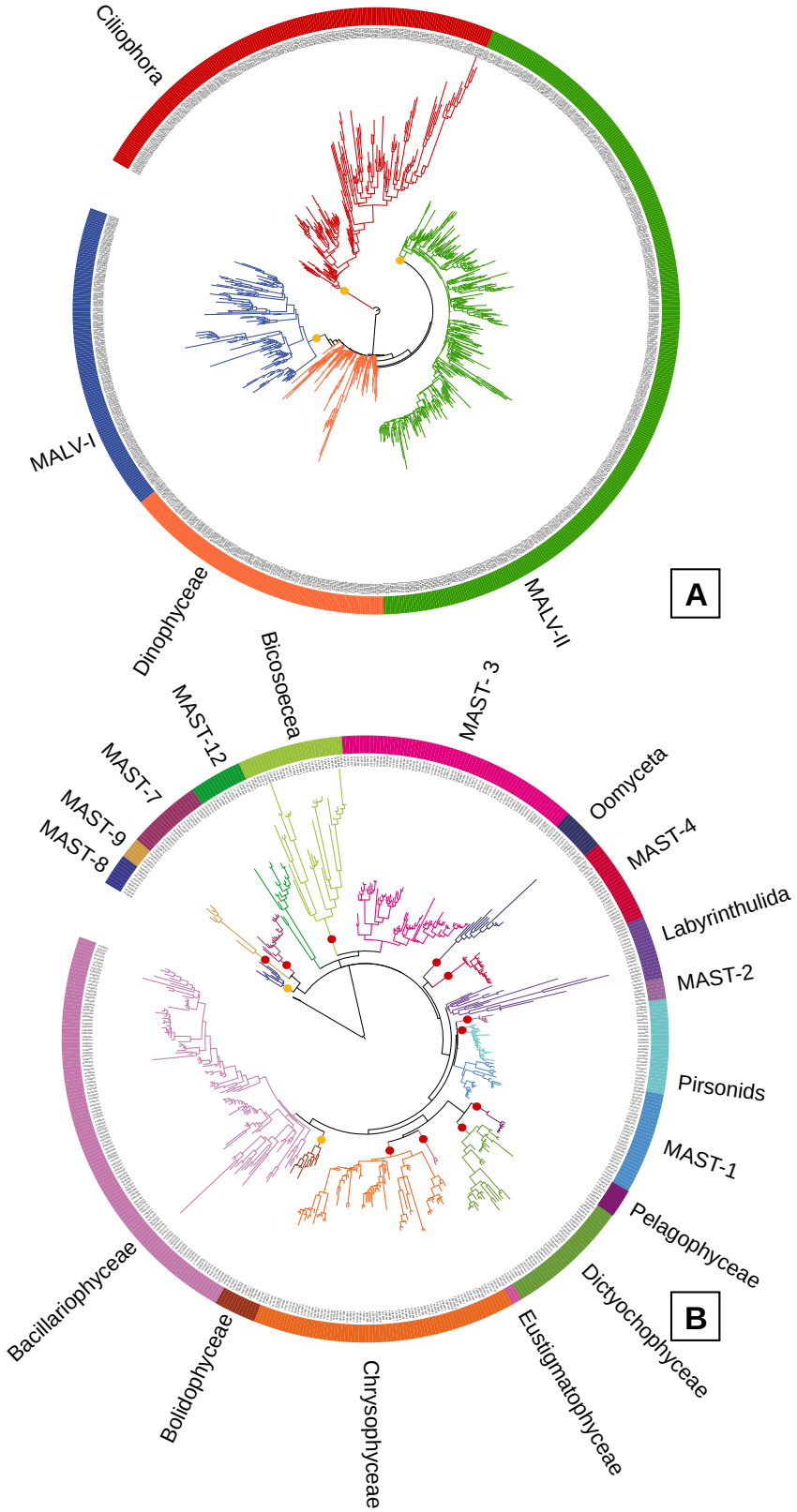


Figure 2. Maximum Likelihood phylogenetic trees for eukaryotic supergroups. Trees include several taxonomic groups within Alveolata (A), Stramenopiles (B), and are done with sequences representative of each OTU obtained clustering at 0.05 distance (A) and 0.01 distance (B). The number of sequences (about 550 bp in length) per tree is 798 and 523 respectively. Red dots represent bootstrap values above 75 and orange dots values above 50.

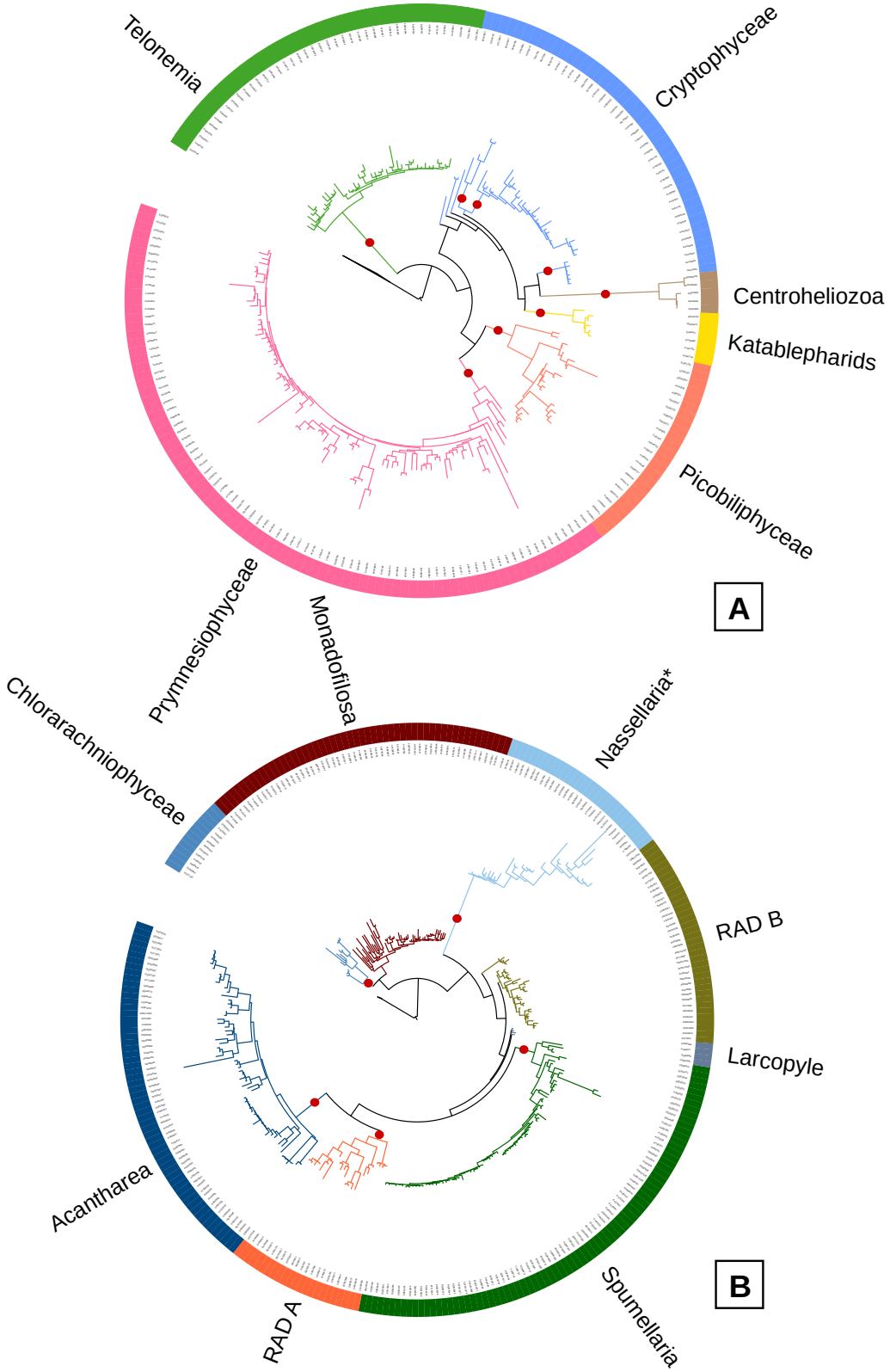


Figure 3. Maximum Likelihood phylogenetic trees for eukaryotic supergroups. Trees include several taxonomic groups within CCTH (A), and Rhizaria (B) and are done with sequences representative of each OTU obtained clustering at 0.05 distance. The number of sequences (about 550 bp in length) per tree 218 and 303 respectively. Red dots represent bootstrap values above 75.

Supergroup phylogenetic trees

Supergroup maximum-likelihood phylogenetic trees were computed to validate the taxonomic assignment of the environmental sequences. The Alveolata tree (Fig. 2A) included only the four largest groups, with one representative sequence from each OTU clustered at 0.05 distance. These groups were well recovered in the tree, but the intragroup topology was not totally correct, since MALV-I and MALV-II emerged from Dinophyceae. Probably the partial region considered (~550 bp) was too short to resolve such a large tree. The other trees were constructed with a representative sequence of each OTU clustered at 0.01 distance. The Stramenopiles tree (Fig. 2B) displayed 18 monophyletic groups, with all photosynthetic groups (Ochrophyta) clustering together. The CCTH tree (Fig. 3A) recovered the monophyly of all groups, except Cryptophyceae. The Rhizaria tree (Fig. 3B) showed the grouping of Chlorarachniophyta and Monadofilosa (from the phylum Cercozoa), while Radiolaria was not well defined as described in previous phylogenies [23]: the class Polycystinea did not appear monophyletic and was separated into the respective orders except Collodaria and Nassellaria that were grouped (as Nassellaria*). These trees confirmed that the final dataset did not contain misclassified sequences. A nexus file of the trees is available as supporting material (Nexus file S1)

Number of OTUs and maximum distance in taxonomic groups

The number of OTUs after clustering sequences at three different cut-off distance levels was estimated for each taxonomic group (Table 1). At 0 distance, the total number of OTUs, calculated for each group and then added up, was 6571. Using the more relaxed criterion of 0.01 distance, to take into account low-frequency sequencing errors and putative intragenomic polymorphisms, resulted in a total count of 3677 OTUs, 2301 of which belonged to Alveolata, 539 to Stramenopiles, 321 to Rhizaria and 213 to CCTH. A substantial decrease of OTUs was observed when clustering at larger distances, with a total number of 1423 OTUs at 0.05 distance.

To report the genetic distance encompassed within groups, we calculated the average, maximum, and maximum corrected pair-wise distances among all sequences within each group (Table 1). The distribution of these values, for the 20 groups having more than 29 sequences, is shown in Fig. S3. The average distance points to the typical distance between any two sequences in a group. It ranged from 0.01 (Pelagophyceae) to 0.23 (Kinetoplastea), with 75% of the cases below 0.14 (Fig. S3). The average distance is a useful descriptor, but it is the maximum distance that defines the group clustering. The intragroup maximum distance ranged from 0.07 (Pelagophyceae) to 0.50 (Dinophyceae), with 75% of the cases below 0.31. The maximum distance, however, could derive from a single highly divergent sequence, which could be fast-evolving or, more critically, could

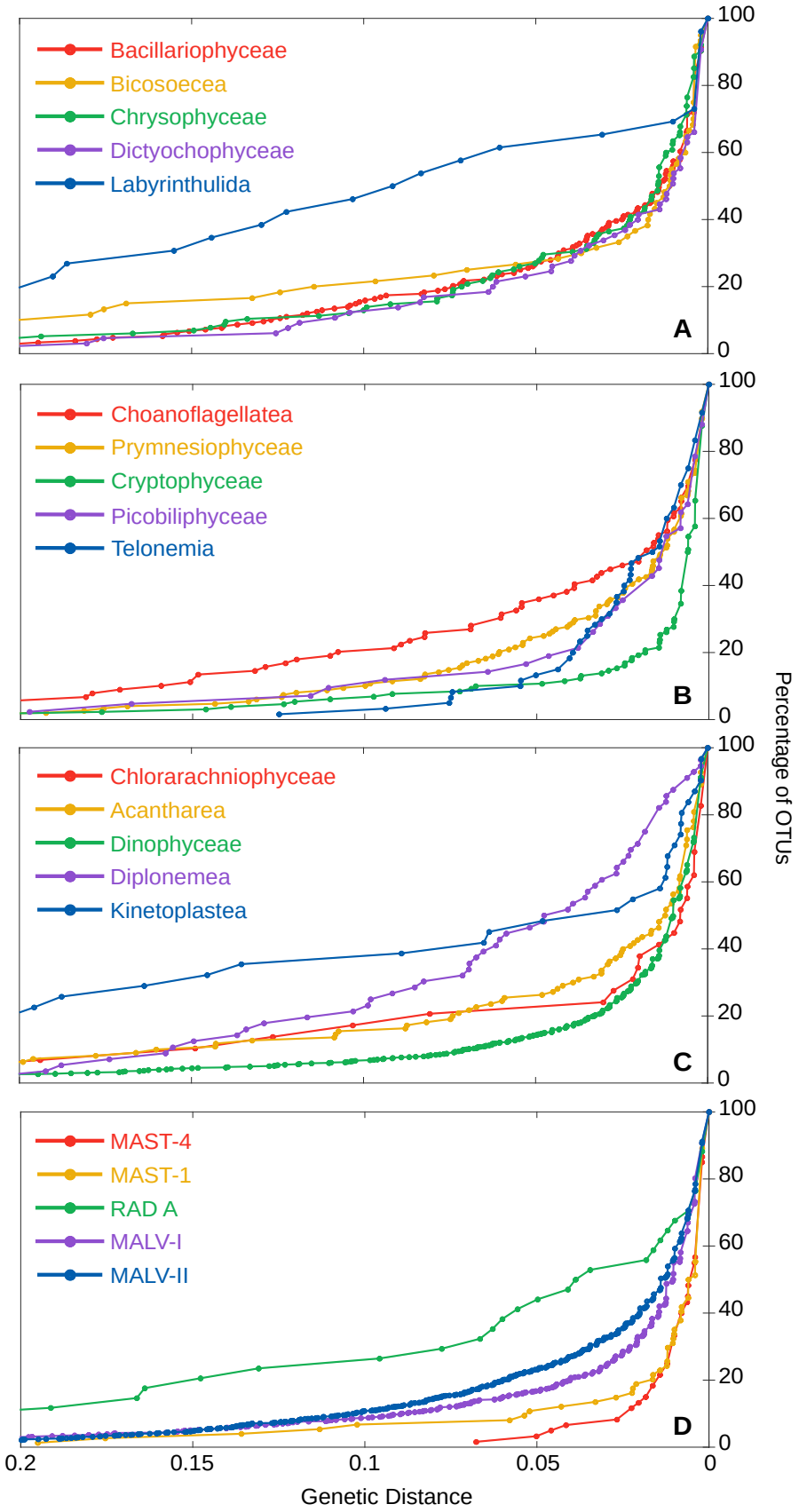


Figure 4. Clustering pattern of several groups of marine protists. The graphs show the percentage of OTUs when sequences are clustered at different genetic distances for several Stramenopiles groups (A), CCTH groups plus Choanoflagellatea (B), Rhizaria and Excavata groups plus Dinophyceae (C) and major ribogroups (D).

contain many sequencing errors. So we proposed another estimate, the maximum corrected distance, as the value at which 90% of sequences cluster in a single OTU. This correction was critical in groups such as Dinophyceae (decrease from 0.50 to 0.24), Prymnesiophyceae, Bolidophyceae or Prasinophyceae, whereas in others the change was minor. Seventy-five percent of the groups exhibited a maximum corrected distance below 0.25. This includes most ribogroups (all MAST clades and RAD B), indicating that these are consistent with taxonomic classes. On the other hand, the maximum corrected distance in MALV-I and MALV-II (0.42 and 0.30, respectively) suggest that these could represent higher taxonomic ranks.

Clustering pattern of taxonomic groups

The clustering pattern was defined as the representation of the number of OTUs obtained in each group when clustering at different cut-off levels (Fig. 4). In order to compare groups, OTU counts were expressed as the percentages of the number detected at 0 distance. A high percentage of OTUs at 0.05 or 0.10 clustering distance would imply the presence of many high-rank lineages. This was the case of Labyrinthulida (Fig. 4A) that showed 65% of OTUs at a distance of 0.05. Similar examples of high-rank diversity were seen in Choanoflagellatea (Fig. 4B), Diplonemea, Kinetoplastea (Fig. 4C) and RAD A (Fig. 4D). In the opposite side of low-rank diversity were the ribogroups MAST-4 and MAST-1 (Fig. 4D), and Cryptophyceae (Fig. 4B) that yielded 2-8% OTUs at a distance of 0.05. Even containing a high number of sequences, the high-rank diversity of Dinophyceae was lower than most other groups.

Phylogenetic structure of taxonomic groups

Lineages Through Time (LTT) plots can be compared using the γ value, which is zero if the rate of cladogenesis was constant through time, negative if it was faster at the origin of the lineage, or positive if it was faster towards the present. Graphically, this is represented by a straight, a concave and a convex line, respectively [14]. The null hypothesis that clades diversified with a constant rate ($\gamma = 0$) was tested with one-tail test, and LLT plots were then displayed per groups that showed γ values significantly negative (Fig. 5A), positive (Fig. 5C) or non-significantly different from zero (Fig. 5B). Labyrinthulida (γ of -3.64) and MALV-II (γ of 16.72) were the two groups with most contrasting patterns, whereas RAD A and Bicosoecea were the ones closest to present a constant rate.

In order to further explore additional features contained in phylogenetic trees, we chose the Stramenopiles supergroup, since all taxonomic groups within this tree appeared monophyletic (Fig. 2B). This was done by using two descriptive parameters: the mean intragroup phylogenetic pair-

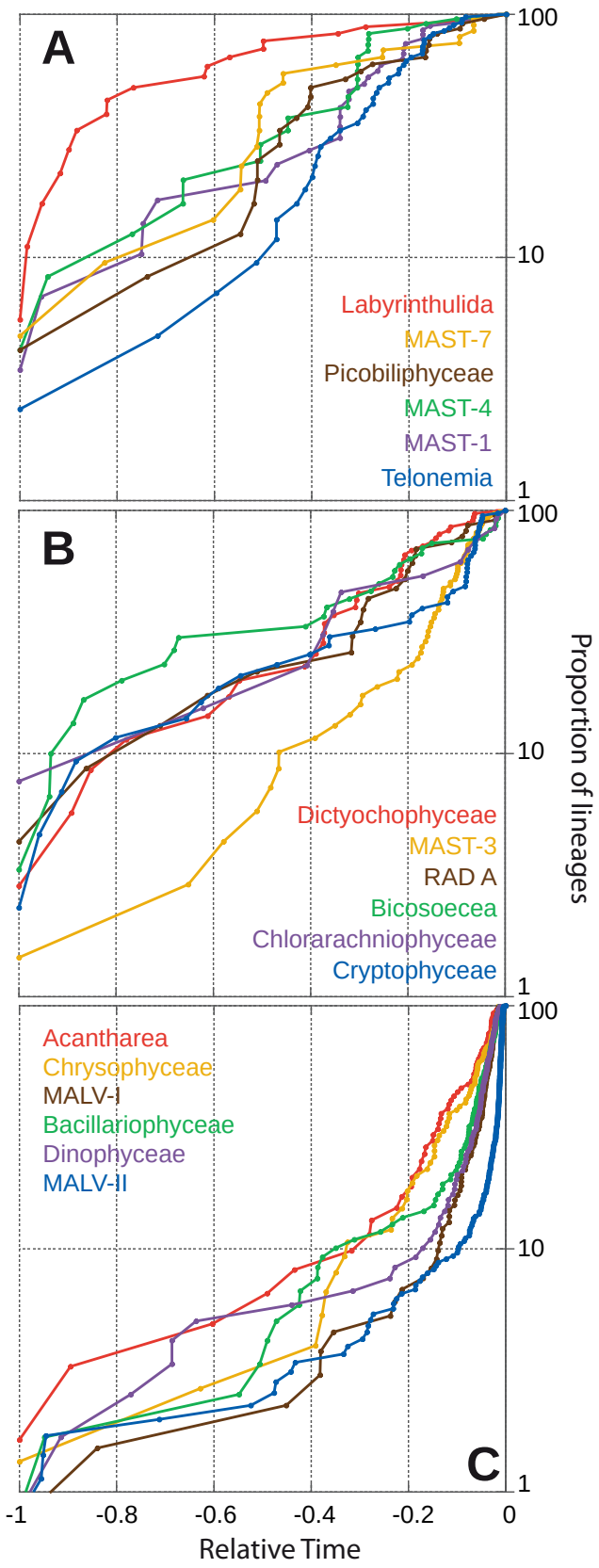


Figure 5. **Phylogenetic structure of several groups of marine protists**. Lineage Through Time (LTT) plots are based on the trees shown in Figure 2-3 and are displayed for groups having $\gamma < 0$ (A), $\gamma = 0$ (B) and $\gamma > 0$ (C), which indicates early, constant or late cladogenesis events, respectively. The number of lineages is standardized to the maximum number at present and relative time is considered.

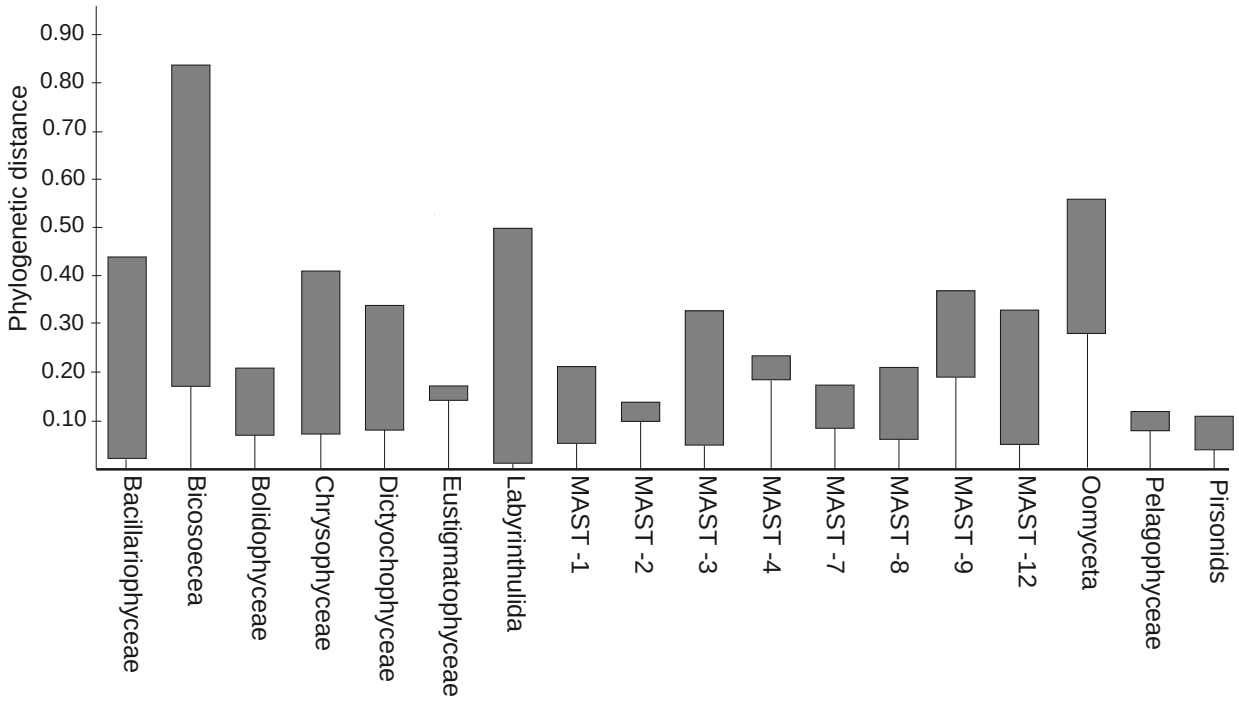


Figure 6. Intragroup phylogenetic distance and trunk length of Stramenopiles groups. A complementary view of phylogenetic structure of Stramenopiles is shown by displaying the trunk length (vertical lines) and the Mean Phylogenetic Distance (vertical boxes) of each group (based on tree in Figure 2B).

wise distance (MPD) and the trunk-length (Fig. 6). There were groups characterized by large intragroup diversity and short trunks, such as Bacillariophyceae and Labyrinthulida, whereas groups like Eustigmatophyceae and MAST-4 presented the opposite structure (short diversity and long trunks). The remaining groups exhibited an intermediate position, some with very high MPD (Bicosoecea, Chrysophyceae and MAST-3) and others with low MPD (MAST-2 and Pelagophyceae). Finally, we generated a matrix of mean distances among sequences belonging to different stramenopiles (Table S3) in order to define the typical distance among groups (including both branch and trunk lengths) and to provide an idea of the phylogenetic differentiation among groups. Bicosoecea was the most isolated lineage, displaying a mean phylogenetic distance of 0.81 to the closest group. On the other hand, the parasitoid group pirsonids was the one exhibiting the lowest distance (0.24) to its closest neighbor.

Discussion

This study is an effort to advance in the understanding of the diversity of marine protists by using publicly available 18S rDNA Sanger environmental sequences. Substantial advances have been gained by sequencing environmental genes using traditional Sanger methods, and the new High Throughput Sequencing (HTS) technologies (e.g. Illumina and 454) are now used to continue

exploring marine microbial diversity [19]. Despite HTS can generate huge amounts of reads from marine microeukaryote communities, we still need a reference frame in order to interpret and organize this flood of new HTS data. Such reference frame, representing the core patterns of marine microeukaryote diversity, needs to be built based on reliable and well curated data. Despite being low-throughput, Sanger sequencing still provides probably the highest quality in sequence data. In addition, Sanger sequences are obtained in a more or less artisanal process that involves, many times, curating carefully each single sequence. For these reasons, we base our analysis in Sanger sequences only.

Our aim was to report for each taxonomic group 1) the number of OTUs and its maximum genetic distance, and 2) the evolutionary patterns inferred from phylogenetic trees. Yet, some preliminary validations were necessary before this analysis. The first step was a proper classification of environmental sequences into classical taxonomic groups or ribogroups. Phylogenetic trees indicated that chimeras or misclassified sequences, which would artificially increase intragroup diversity, were accurately removed. The second step was identifying a useful 18S rDNA region. The V4-V5 hypervariable region, widely used in environmental surveys [24,25], provided accurate phylogenies and resulted to be a good descriptor of the variability of the entire 18S rRNA gene, overestimating pairwise distances by a factor of ~1.4. The V9 region, optimal for early pyrosequencing technologies due to its short size [19,26], was already known to lack specific signatures for higher-level taxa [27], and in our analysis was a poor predictor of the whole gene variability. Similar results had been obtained when comparing complete 18S rDNA and V9 regions [28] although with a lower coefficient ($R^2=0.40$) and higher slope ($m=1.86$), probably because this study did not perform a separate analysis per supergroup as we did here. The third step was to find out specific clustering cut-off levels that define taxonomic ranks. While some studies have investigated the level corresponding to the rank species [15,16], very little has been done for higher rank categories. Regarding the clustering at the class level, 75% of the groups had a maximum corrected distance (at the V4-V5 region) below 0.25 (the full gene distance could be grossly calculated by dividing times 1.4). This was the general picture, since evolutionary rates might differ among slow- and fast-evolving lineages. Remarkably, many of the arbitrarily defined environmental ribogroups (MALV-III, MALV-V, RAD B and all MAST clades) were consistent with this maximum distance, indicating that they were congruent with a taxonomic rank equivalent to the classical class.

Once the dataset was manually curated and all sequences assigned to one of the 65 taxonomic groups, we started to analyze the diversity of the whole dataset of marine microeukaryotes. Overall, we detected 3,677 OTUs at 0.01 distance, mostly within Alveolata (63% of OTUs), Stramenopiles (15%), Rhizaria (9%) and CCTH (6%). Almost half of these OTUs belonged to taxonomically

undefined ribogroups. The poor representation of the supergroups Amoebozoa and Excavata probably reflects their lower relative abundance as compared with the other supergroups in the marine plankton. This taxonomic distribution was similar to previously reviewed data [2] and could be influenced by methodological biases affecting the real proportion of taxa in natural samples. Since sequences came from libraries prepared from extracted DNA, some could derive from non-living or non-active organisms [4,10], and taxa with high rDNA copy number could be overrepresented [29]. The moderate levels of diversity observed here were lower than what has been observed in seminal pyrosequencing studies [28,30]. Even the groups with more sequences did not saturate, and rarefaction curves never reached a plateau (data not shown). Despite the dataset analyzed here most likely captures the general architecture of protist diversity in terms of main phylogenetic lineages, it is clear that a better estimation of diversity extent requires deeper sequencing efforts as provided by HTS. When observing how the clustering threshold affected OTU numbers, Alveolata still dominated at all levels, whereas classes like Labyrinthulida, Diplonemea and Kinetoplastea had an exceptionally high diversity. The last one exhibited the highest maximum corrected distance, probably due to a massive accumulation of sequence mutations [31].

Whereas the clustering pattern (Fig. 4) allowed quantifying the degree of genetic diversity of the groups at present time, the LTT plots (Fig. 5 and Fig. S4) used the tree topology to infer the cladogenesis events during the entire evolutionary history of different groups. It should be noted that incomplete taxon sampling could lead to the incorrect conclusion that speciation and extinction rates varied through time [32]. Other phenomena may give the false impression of non-constant rate of cladogenesis. Thus, the fact that only clades that survived to the present are considered may result in higher apparent rate of cladogenesis at the beginning of the lineage (a phenomenon known as «push of the past»), whereas higher rate of cladogenesis towards the present may be because lineages arising in recent times have had less time to go extinct («pull of the present») [33]. Overall, the trend of cladogenesis through time is well described by the γ value [14]. The expected tendency is to find early cladogenesis events followed by a slowdown towards the present, with γ values below 0, as commonly seen in animals and plants [34]. However, microorganisms, with their huge populations sizes (and likely lower extinction rates), may deviate from this general trend. Preliminary data showed that microbial eukaryotes had negative γ whereas prokaryotes tended to have a constant rate [14], or an increase in cladogenesis towards the present [35], although this latter trend could partly be due to the pull of the present phenomenon. Our results illustrated three evolutionary scenarios, with microeukaryote groups exhibiting early, constant, or late cladogenesis events. Thus, both Labyrinthulida and MAST-4 had early cladogenesis, even though Labyrinthulida was more diverse, perhaps because it was an early-diverging lineage [36].

Remarkably, half of the groups from our study had a positive γ (MALV-II showed the highest value), therefore deviating from the general pattern for plants and animals.

Phylogenetic supergroup trees displayed a branch distance that was not used in LTT plots, the trunk at the base of each monophyletic group. The trunk length represents the evolutionary time between the first appearance of the group and its observed diversification (putative diversifying lineages during this time are extinct). In a complete phylogeny, this trunk is a key feature to understand the intergroup diversity and complements the information given by MPD (Mean Phylogenetic Distance). Using the Stramenopiles tree as model for this analysis, it became evident that the MPD was not enough to describe the genetic isolation of a group, as confirmed by the minimum intergroup distance (Table S2). For instance, the Oomyceta had a lower MPD than Labyrinthulida and Bacillariophyceae, but a larger minimum distance (and trunk length) with its closer neighbor.

In summary, a good approximation to the evolutionary history of a given group could be reached by combining LTT plots and trunk lengths. This provided an overview of when most diversification occurred and what was the uniqueness of each group. The phylogenetic structure enriched and complemented the picture drawn by clustering pattern, which allowed reasonable comparisons among groups in terms of OTU numbers and maximum distances. Together, these two structural features gave a reasonable characterization of the diversity of the main microeukaryote clades. New sequencing technologies (pyrosequencing, Illumina) are already providing a huge amount of sequences, and a good phylogenetic and clustering pattern overview based on a robust technique is required to ensure a solid backbone for interpreting and manipulating future high-throughput datasets.

Materials and Methods

Sequence dataset and classification into taxonomic groups

The initial set of 163,975 sequences derived from molecular surveys of 18S rDNA genes published in GenBank until January 2010 (see Table S1) plus a few (<5%) unpublished sequences obtained at the Station Biologique de Roscoff (France). The database was filtered to keep sequences longer than 500 bp from marine planktonic protists (excluding sequences retrieved in freshwaters and sediments, or affiliating to metazoans and fungi). In addition, the sequence quality of the dataset was refined by keeping only sequences derived from clone libraries, having few unidentified bases (if any), and that passed a chimera check done with the application KeyDNATools (Fig. S1).

The resultant 13,270 sequences were taxonomically classified with KeyDNATools (Fig. S2). Sequences ambiguously classified (less than 5 keys, keys in one region of the sequence only, or few keys from different groups [non-obvious chimeras]) were checked with BLAST [37] and assigned to a given group if they were ≥90% similar to a well-identified reference sequence. In some cases, BLAST with different parts of the sequence was done to double-check they were not chimeras. The initial dataset was distributed into 65 taxonomic groups (basically based in the «Second rank» level of Adl et al. [38]), including classical taxa mostly at the «Class» level plus new ribogroups. Sequences within each group were aligned with the FFT-NS-i strategy of MAFFT [39]. The alignment was cut manually in Seaview 3.2 [40] to keep a dataset of ~500 bp that covered the V4-V5 regions of the 18S rDNA. Sequences shorter than 475 bp were eliminated. This process resulted in 8291 well-identified sequences plus a miscellaneous assemblage of 427 sequences that could not be placed in any taxonomic group (named Novel). A fasta file with all sequences and a text file with their affiliation are available from the authors upon request.

Comparing different regions of the 18S rDNA

Full-length 18S rDNA sequences were prepared from three major supergroups: Rhizaria (72 sequences), Stramenopiles (60 sequences) and Alveolata (232 sequences). These were aligned with MAFFT as before and two regional alignments were extracted from the full gene alignments. The V4-V5 region was composed by the V4 region delimited by primers TAReuk454FWD1 (5'-CCAGCA(G/C)C(C/T)GCGGTAATTCC-3', *S. cerevisiae* [U53879] positions 565-584) and TAReukREV3 (5'-ACTTTCGTTCTTGAT(C/T)(A/G)A-3', positions 964-981) [19] and the following ~100 bp forming the V5 region. The V9 region was delimited by primers 1391F (5'-GTACACACCGCCCGTC-3', positions 1629-1644), and EukB (5'-TGATCCTTCTGCAGGTTCACCTAC-3', positions 1774-1797). The V4 forward and V9 reverse primers were excluded from the alignments.

Distance estimates and sequence clustering

Sequence alignments were processed with PAUP [41] to generate a pair-wise genetic distance matrix with Jukes-Cantor as the substitution model. The matrix was used to calculate the average distance within a group (the mean of all pair-wise distances) and also its maximum distance (the highest pair-wise distance value). The distance matrix was also used to cluster sequences in OTUs (Operational Taxonomic Units) at different distance levels with MOTHUR [42], with default settings of furthest neighbor and maximum precision (precision=10,000). This clustering routine was also used to calculate a third estimate for each group (maximum corrected distance), which was defined as the distance at which 90% of the sequences cluster to form a single OTU.

Phylogenetic analysis

Phylogenetic trees were constructed using one representative sequence from each OTU, generated using a clustering threshold of 0.01 (Stramenopiles, Rhizaria and CCTH) or 0.05 (Alveolata). OTU clustering was done separately for each taxonomic group, then representative sequences from the same supergroup were combined and aligned with MAFFT. Maximum-likelihood phylogenetic trees were done with RAxML [43] at the University of Oslo Bioportal (www.bioportal.uio.no), using the GTR-GAMMA evolutionary model and performing 100 alternative searches for topology and bootstrap using distinct random starting trees. Phylogenetic trees were visualized with the online tool iTOL [44]. Supergroup trees are available from the authors upon request.

For each taxonomic group within Stramenopiles, the mean phylogenetic distance (MPD) was calculated with PHYLOCOM [45]. This software was also used to estimate the length of the branch at the base of each monophyletic group, which was named «trunk», and the average inter-group phylogenetic distance (the mean of all pair-wise distances between sequences from different groups). Phylogenetic trees representing the different taxonomic groups were extracted from the Stramenopiles tree using Dendroscope [46]. Trees were transformed to ultrametric, and used to calculate the evolution of the lineages through time (LTT). Relative time was considered, ranging from -1 (the origin of the lineage) to 0 (present time), and the number of lineages was standardized (percentage of the maximum number) to compare LTT plots among groups. For each plot, the γ-statistic was calculated as a descriptor of the evolutionary trends [32]. All analyses were carried in R environment (http://www.r-project.org/) using APE [47] and LASER [48] packages.

Acknowledgements

Funding has been provided by projects FLAME (CGL2010-16304, MICINN, Spain) and BioMarKs (2008-6530, ERA-net Biodiversa, EU) to R.M., by project GEMMA (CTM2007-63753-C02-01/MAR, MEC, Spain) to Carlos Pedrós-Alió, by a F.P.I. fellowship from the Spanish Ministry of Education and Science to M.C.P., and by a Marie Curie Intra-European Fellowship grant PIEF-GA-2009-235365 to R.L. Original sequence database was built under the frame of the French ANR-Biodiversité project AQUAPARADOX.

References

1. Epstein S, López-García P (2008) "Missing" protists: a molecular prospective. Biodivers Conserv 17: 261-276.

2. Massana R, Pedrós-Alió C (2008) Unveiling new microbial eukaryotes in the surface ocean. Curr Opin Microbiol 11: 213-218.

3. Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton (≤3 μm) in marine ecosystems. FEMS Microbiol Rev 32: 795-820.

4. Not F, del Campo J, Balagué V, de Vargas C, Massana R (2009) New Insights into the Diversity of Marine Picoeukaryotes. PLoS ONE 4: e7143.

5. Savin MC, Martin JL, LeGresley M, Giewat M, Rooney-Varga J (2004) Plankton Diversity in the Bay of Fundy as Measured by Morphological and Molecular Methods. Microb Ecol 48: 51-65.

6. Terrado R, Vincent W, Lovejoy C (2009) Mesopelagic protists: diversity and succession in a coastal Arctic ecosystem. Aquat Microb Ecol 56: 25-40.

7. Jeon S, Bunge J, Leslin C, Stoeck T, Hong S, et al. (2008) Environmental rRNA inventories miss over half of protistan diversity. BMC Microbiol 8: 222.

8. Massana R, Pernice M, Bunge JA, Campo Jd (2011) Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean. ISME J 5: 184-195.

9. Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ (2007) Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. Environ Microbiol 9: 1233-1252.

10. Stoeck T, Kasper J, Bunge J, Leslin C, Ilyin V, et al. (2007) Protistan Diversity in the Arctic: A Case of Paleoclimate Shaping Modern Biodiversity? PLoS ONE 2: e728.

11. Kirkpatrick M, Slatkin M (1993) Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. Evolution 47: 1171-1181.

12. Vamosi SM, Heard SB, Vamosi JC, Webb CO (2009) Emerging patterns in the comparative analysis of phylogenetic community structure. Mol Ecol 18: 572-592.

13. Purvis A, Nee S, Harvey PH (1995) Macroevolutionary Inferences from Primate Phylogeny. Proceedings of the Royal Society of London Series B: Biol Sci 260: 329-333.

14. Martin AP, Costello EK, Meyer AF, Nemergut DR, Schmidt SK, et al. (2004) The rate and pattern of cladogenesis in microbes. Evolution 58: 946-955.

15. Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, et al. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. Appl Environ Microbiol 75: 5797-5808.

16. Nebel M, Pfabel C, Stock A, Dunthorn M, Stoeck T (2011) Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. Environ Microbiol Reports 3: 154-158.

17. Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, et al. (2008) Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environ Microbiol 10: 397-408.

18. Massana R, Castresana J, Balague V, Guillou L, Romari K, et al. (2004) Phylogenetic and ecological analysis of novel marine stramenopiles. Appl Environ Microbiol 70: 3528-3534.

19. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, et al. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19: 21-31.

20. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12: 118-123.

21. Quince (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Meth 6: 639-641.

22. Burki F, Inagaki Y, Bråte J, Archibald JM, Keeling PJ, et al. (2009) Large-Scale Phylogenomic Analyses Reveal That Two Enigmatic Protist Lineages, Telonemia and Centroheliozoa, Are Related to Photosynthetic Chromalveolates. Gen Biol Evol 1: 231-238.

23. Krabberød AK, Bråte J, Dolven JK, Ose RF, Klaveness D, et al. (2011) Radiolaria Divided into Polycystina and Spasmaria in Combined 18S and 28S rDNA Phylogeny. PLoS ONE 6: e23526.

24. Brate J, Logares R, Berney C, Ree DK, Klaveness D, et al. (2010) Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environ-

mental rDNA. ISME J 4: 1144-1153.

25. Dunthorn M, Klier J, Bunge J, Stoeck T (2012) Comparing the Hyper-Variable V4 and V9 Regions of the Small Subunit rDNA for Assessment of Ciliate Environmental Diversity. J Eukaryot Microbiol 59: 185-187.

26. Behnke A, Engel M, Christen R, Nebel M, Klein RR, et al. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environ Microbiol 13: 340-349.

27. Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, et al. (2011) Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. PLoS ONE 6: e18169.

28. Brown MV, Philip GK, Bunge JA, Smith MC, Bisset A, et al. (2009) Microbial community structure in the North Pacific Ocean. ISME J 3: 1374-1386.

29. Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol Ecol 52: 79-92.

30. Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK (2010) Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. ISME J 4: 1053-1059.

31. Moreira D, López-García P, Vickerman K (2004) An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. Int J Syst Evol Microbiol 54: 1861-1875.

32. Pybus OG, Harvey PH (2000) Testing Macro-Evolutionary Models Using Incomplete Molecular Phylogenies. Proceedings: Bio Sci 267: 2267-2272.

33. Nee S, Holmes EC, May RM, Harvey PH (1994) Extinction Rates can be Estimated from Molecular Phylogenies. Philosophical Transactions of the Royal Society of London Series B: Biol Sci 344: 77-82.

34. McPeek Mark A (2008) The Ecological Dynamics of Clade Diversification and Community Assembly. Am Nat 172: E270-E284.

35. Barberán A, Fernández-Guerra A, Auguet J-C, Galand PE, Casamayor EO (2011) Phylogenetic ecology of widespread uncultured clades of the Kingdom Euryarchaeota. Mol Ecol

20: 1988-1996.

36. Riisberg I, Orr RJS, Kluge R, Shalchian-Tabrizi K, Bowers HA, et al. (2009) Seven Gene Phylogeny of Heterokonts. Protist 160: 191-204.

37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

38. Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, et al. (2005) The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists. J Eukaryot Microbiol 52: 399-451.

39. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

40. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12: 543-548.

41. Swofford D (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.

42. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75: 7537-7541.

43. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

44. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.

45. Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24: 2098-2100.

46. Huson D, Richter D, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.

47. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290.

48. Rabosky DL (2006) LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evolutionary bioinformatics online 2: 247-250.

Table S1. List of all studies from which we have retrieved the 18S rDNA environmental sequences.

Year	First author	Reference	Size fraction (µm)
2001	Díez	Appl Environ Microbiol **68**:4554-4558	0.2 - 2
2001	López-García	Nature **409**:603-607	0.2 - 5
2001	Moon-van der Staay	Nature **409**:607-610	0.2 - 3
2003	Stoeck	Appl Environ Microbiol **69**:5656-5663	Whole
2004	Corredor	Appl Environ Microbiol **70**:5459-5468	Whole
2004	Massana	Appl Environ Microbiol **70**:3528-3534	0.2 -3
2004	Romari	Limnol Oceanogr **49**: 784-798	0.2 - 3
2004	Savin	Microb Ecol **48**: 51-65	5 - 100
2004	Yuan	FEMS Microbiol Let **240**: 163-170	Whole
2006	Behnke	Appl Environ Microbiol **72**:3626-3636	Whole
2006	Lovejoy	Appl Environ Microbiol **72**:3085-3095	0.2 - 3
2006	Massana	Aquat Microb Ecol **45**:171-180	0.2 - 3
2006	Medlin	Microb Ecol **52**: 53-71	0.2 - 3
2006	Stoeck	Protist **157**: 31-43	Whole
2006	Worden	Aquat Microb Ecol **43**:165-175	0.45 - 2
2006	Zuendorf	FEMS Microbiol Ecol **58**: 476-491	Whole
2007	Bass	Proc Roy Soc Lond B **274**: 3069-3077	Whole
2007	Countway	Environ Microbiol **9**: 1219-1232	0.2 - 200
2007	López-García	Environ Microbiol **9**: 546-554	Whole
2007	Massana	Environ Microbiol **9**: 2260-2269	0.2 - 3
2007	Not	Environ Microbiol **9**: 1233-1252	0.2 - 2
2007	Stoeck	Microb Ecol **53**: 328-339	Whole
2008	Amaral-Zettler	Environ Sci Technol **42**: 9072-9080	Whole
2008	Guillou	Environ Microbiol **10**:3349-3365	Various
2008	Not	Deep Sea Res Part I **55**: 1456-1473	0.2 - 3
2009	Alexander	Environ Microbiol **11**: 360-381	Whole
2009	Amacher	Deep Sea Res Part I **56**: 2206-2215	Whole
2009	Caron	Appl Environ Microbiol **75**:5797-5808	0.2 - 200
2009	Luo	Hydrobiologia **636**: 233-248	0.2 – 50 / Whole
2009	Not	PLoS ONE **4**:e7143	0.6 - 3
2009	Potvin	J Eukaryot Microbiol **56**: 174-181	0.2 - 3
2009	Shi	PLoS ONE **4**:e7657	0.2 - 3
2009	Terrado	Aquat Microb Ecol **56**:25-39	0.2 – 3 / 3 - Whole
2010	del Campo	Prot.**162**: 435-448	0.2 - 3

Table S2. Classification of environmental 18S rDNA sequences in 23 taxonomic groups.

Supergroup	Group		Distances				OTUs		
			Seq	Avg	Max	Max_c	100%	99%	95%
Amoebozoa	*Breviata*	G	3	0.19	0.27	-	3	**3**	3
	Lobosa	D	8	0.31	0.52	-	8	**7**	5
	Other Amoebozoa	-	1	-	-	-	1	**1**	1
	Ichthyosporea	C	1	-	-	-	1	**1**	1
Rhizaria	*Endomyxa*	S	3	0.27	0.30	-	3	**3**	3
	Other Cercozoa	-	31	-	-	-	23	**15**	12
Archaeplastida	*Chlorophyceae*	C	5	0.08	0.13	-	5	**5**	4
	Embryophyceae	C	6	0.13	0.26	-	6	**3**	3
	Florideophyceae	C	1	-	-	-	1	**1**	1
	Ulvophyceae	C	1	-	-	-	1	**1**	1
Stramenopiles	*MAST-9*	R	8	0.08	0.17	-	8	**6**	4
	Phaeophyceae	C	3	0.02	0.03	-	3	**3**	1
	Planomonadida	C	1	-	-	-	1	**1**	1
	Raphidophyceae	C	2	-	0.01	-	2	**1**	1
	Xanthophyceae	C	1	-	-	-	1	**1**	1
CCTH	*Centroheliozoa*	C	8	0.06	0.11	-	7	**5**	2
	Pavlovophyceae	C	1	-	-	-	1	**1**	1
Alveolata	*Apicomplexa*	C	6	0.28	0.42	-	6	**6**	6
	Ellobiopsidae	C	1	-	-	-	1	**1**	1
	MALV-IV	R	5	0.05	0.09	-	5	**5**	3
	Perkinsea	C	4	0.13	0.16	-	4	**4**	4
Excavata	*Eopharyngia*	C	3	0.00	0.00	-	3	**1**	1
	Jacobea	C	6	0.20	0.44	-	6	**4**	3

In this table are shown groups with less than 10 sequences. The groups are coded according to their taxonomic rank (D: division; P: phylum; S: subphylum; C: class; G: genus; R: ribogroup). The table shows the number of sequences per group (Seq), the average (Avg), maximum (Max) and maximum corrected (Max_c) pair-wise distances, and the number of OTUs at three cut-off levels.

Table S3. Matrix of mean distances among sequences belonging to different stramenopiles. In bold there is the minimum distance between groups

	Bicosoecida	Bolidophyceae	Chrysophyceae	Dictyophyceae	Eustigmatophyceae	Labyrinthulidae	MAST1	MAST2	MAST3	MAST4	MAST7	MAST8	MAST9	MAST12	Oomyceta	Pelagophyceae	Pirsonia
Bacillariophyta	1.05	0.47	0.63	0.7	0.52	0.65	0.49	0.48	0.68	0.65	0.69	0.6	0.92	0.85	0.83	0.6	**0.45**
Bicosoecida	0	0.94	1.04	1.11	0.93	0.97	0.86	0.85	**0.81**	0.91	0.92	0.84	1.16	1.02	1.08	1.01	0.82
Bolidophyceae	0.94	0	0.52	0.59	0.41	0.53	0.38	0.36	0.57	0.54	0.57	0.48	0.81	0.74	0.71	0.49	**0.34**
Chrysophyceae	1.04	0.52	0	0.62	**0.42**	0.63	0.48	0.47	0.67	0.64	0.68	0.59	0.91	0.84	0.81	0.52	0.44
Dictyophyceae	1.11	0.59	0.62	0	0.52	0.71	0.56	0.54	0.75	0.71	0.75	0.66	0.98	0.91	0.89	**0.42**	0.51
Eustigmatophyceae	0.93	0.41	0.42	0.52	0	0.53	0.38	0.36	0.57	0.53	0.57	0.48	0.8	0.73	0.71	0.42	**0.33**
Labyrinthulidae	0.97	0.53	0.63	0.71	0.53	0	0.46	0.44	0.61	0.57	0.61	0.52	0.84	0.78	0.75	0.61	**0.42**
MAST1	0.86	0.38	0.48	0.56	0.38	0.46	0	0.28	0.5	0.47	0.5	0.41	0.73	0.67	0.64	0.46	**0.25**
MAST2	0.85	0.36	0.47	0.54	0.36	0.44	0.28	0	0.48	0.45	0.49	0.4	0.72	0.65	0.62	0.44	**0.24**
MAST3	0.81	0.57	0.67	0.75	0.57	0.61	0.5	0.48	0	0.54	0.56	0.47	0.79	0.65	0.72	0.65	**0.46**
MAST4	0.91	0.54	0.64	0.71	0.53	0.57	0.47	0.45	0.54	0	0.55	0.46	0.78	0.71	0.62	0.61	**0.42**
MAST7	0.92	0.57	0.68	0.75	0.57	0.61	0.5	0.49	0.56	0.55	0	**0.38**	0.61	0.73	0.72	0.65	0.46
MAST8	0.84	0.48	0.59	0.66	0.48	0.52	0.41	0.4	0.47	0.46	0.38	0	0.62	0.64	0.63	0.56	**0.37**
MAST9	1.16	0.81	0.91	0.98	0.8	0.84	0.73	0.72	0.79	0.78	**0.61**	0.62	0	0.96	0.95	0.88	0.69
MAST12	1.02	0.74	0.84	0.91	0.73	0.78	0.67	0.65	0.65	0.71	0.73	0.64	0.96	0	0.88	0.81	**0.62**
Oomyceta	1.08	0.71	0.81	0.89	0.71	0.75	0.64	0.62	0.72	0.62	0.72	0.63	0.95	0.88	0	0.79	**0.6**
Pelagophyceae	1.01	0.49	0.52	0.42	0.42	0.61	0.46	0.44	0.65	0.61	0.65	0.56	0.88	0.81	0.79	0	**0.41**
Pirsonia	0.82	0.34	0.44	0.51	0.33	0.42	0.25	**0.24**	0.46	0.42	0.46	0.37	0.69	0.62	0.6	0.41	0

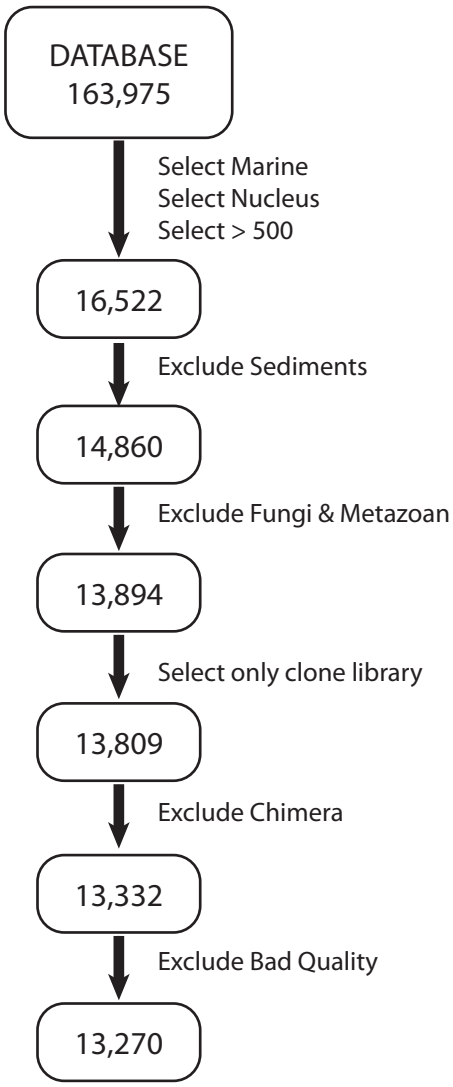


Figure S1. Pipeline for database treatment. Processing of environmental 18S rDNA sequences from initial database to working dataset, showing the number of sequences left after each filtering step.

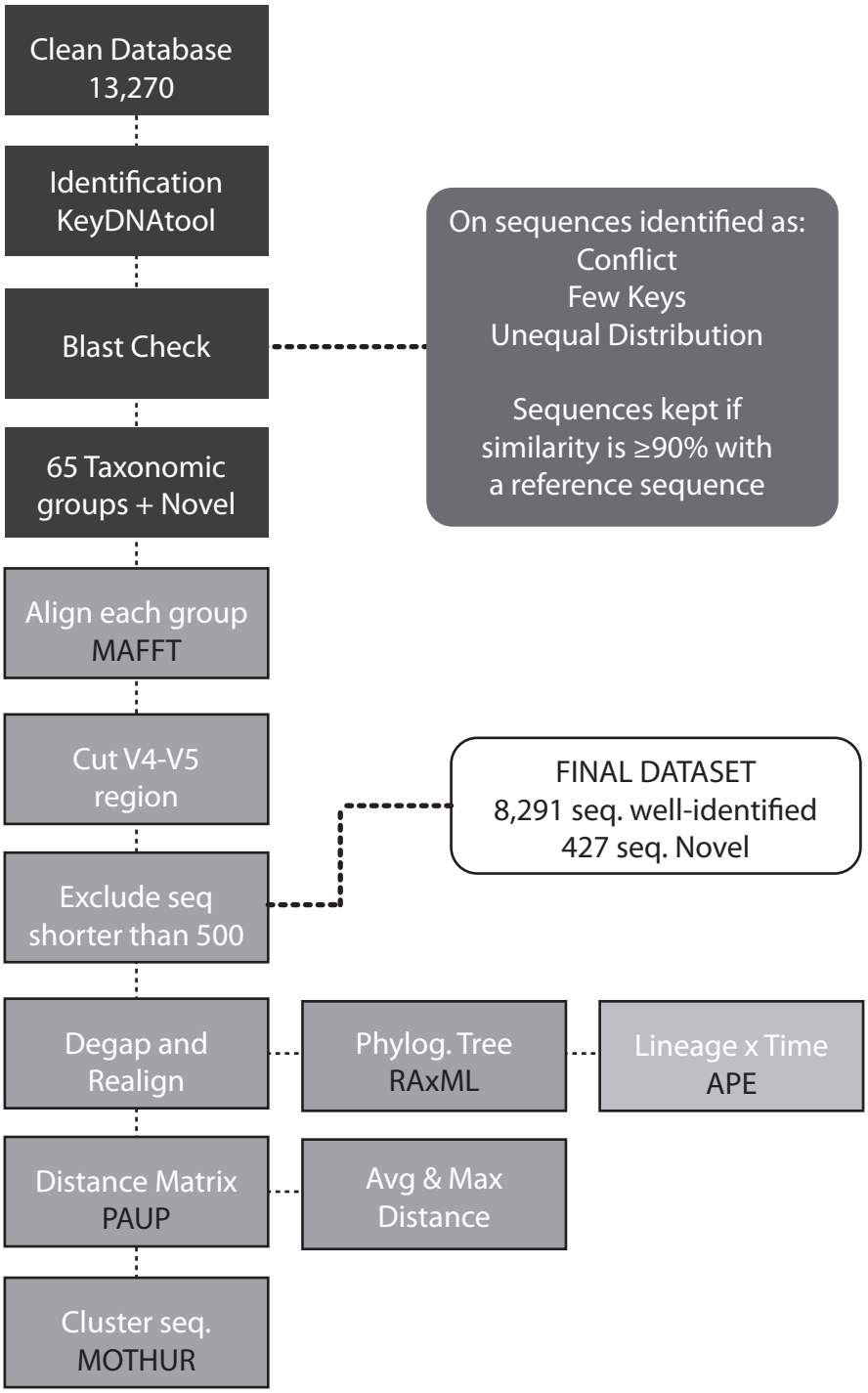


Figure S2. Pipeline for sequence treatment. Dark grey boxes are analyses performed on the entire dataset to split sequences into 65 taxonomic groups (plus the unassigned sequences as "Novel"). Light grey boxes are analyses performed on each of the 65 groups.

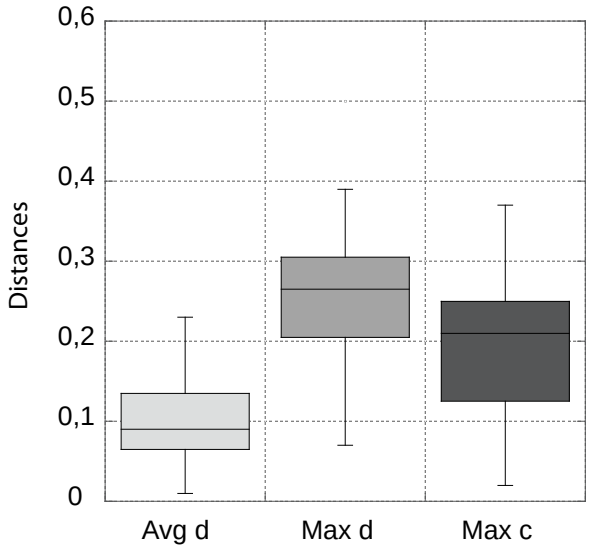


Figure S3. Genetic distances. Distribution of Average, Maximum and Maximum corrected distances within the 20 classes that have more than 30 sequences.

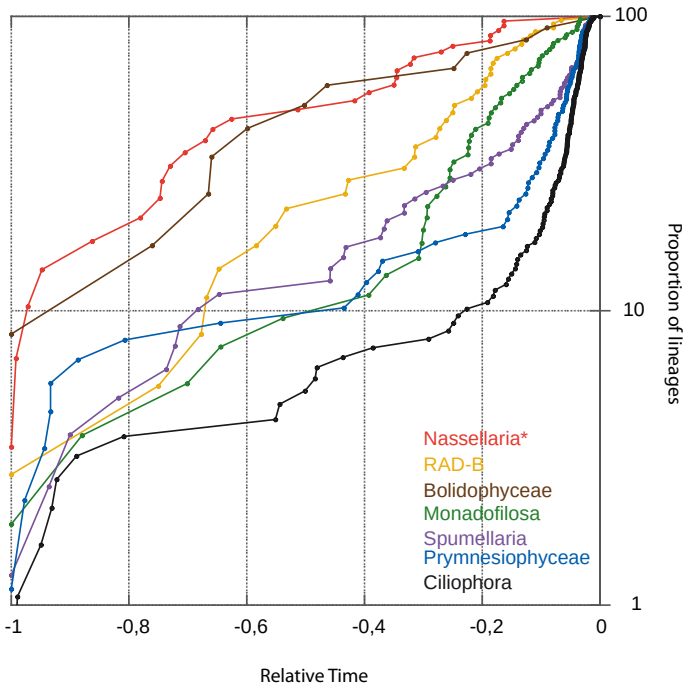


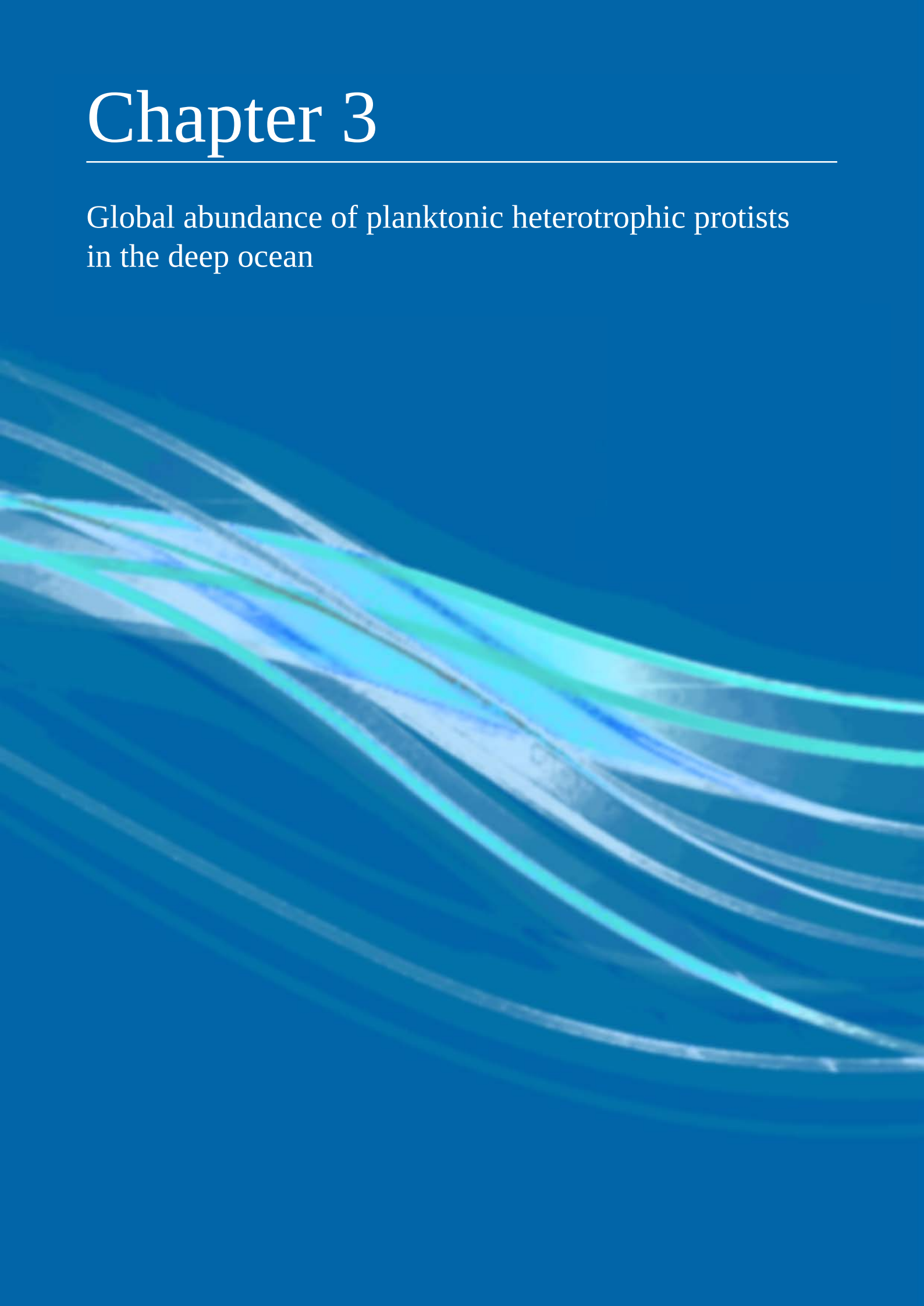
Figure S4. Phylogenetic structure of several groups of marine protists. Lineage Through Time (LTT) plots are based on the trees shown in Figure 2-3 and are displayed for groups having $\gamma < 0$ (Nassellaria-Collodaria, RAD B), $\gamma = 0$ (Bolidophyceae, Monadofilosa) and $\gamma > 0$ (Spumellaria, Prymnesiophyceae, Ciliophora), which indicates early, constant or late cladogenesis events, respectively. The number of lineages is standardized to the maximum number at present and relative time is considered.

Cambia lo superficial
Cambia también lo profundo
Cambia el modo de pensar
Cambia todo en este mundo

Julio Numhauser,
Todo cambia

Chapter 3

Global abundance of planktonic heterotrophic protists in the deep ocean



Pernice MC, Forn I, Gomes A, Lara E, Vaqué D, Duarte CM, Gasol JM, Massana R. Global Abundance of planktonic heterotrophic protists in the deep ocean. Submitted to *The ISME Journal*.

Abstract

The dark ocean is one of the largest biomes on Earth, with critical roles in organic matter remineralization and global carbon sequestration. Despite its recognized importance, little is known about some key microbial players such as the community of heterotrophic protists (HP), which are likely the main consumers of prokaryotic biomass. To investigate this microbial component at a global scale, we determined their abundance and biomass in meso- and bathypelagic waters in samples from the Malaspina-2010 circumnavigation using a combination of epifluorescence microscopy and flow cytometry. HP were clearly ubiquitous in the global deep ocean, even at the deepest 4000 m samples investigated. Their abundances decreased with depth, from an average of 72 ± 19 cells mL^{-1} in mesopelagic waters to 11 ± 1 cells mL^{-1} in bathypelagic waters, whereas their global biomass decreased from 280 ± 46 to 50 ± 14 pg C mL^{-1}. The parameters that better explained the variance of HP abundance were depth and prokaryote abundance, and to lesser extent oxygen and Large Viruses. Different signs suggested active grazing of HP on prokaryotes, such as the presence of flagella in most cells, and the generally good correlation with prokaryote abundance. On a finer scale, the prokaryote:HP ratio in abundance varied at a regional scale, and sites with the highest ratios appear related to a larger contribution of osmotrophy. Our study allows a better understanding of the relation between HP and their environment, shedding light onto their importance as players in the dark ocean's microbial food web.

Introduction

The peculiar aphotic property of the so-called dark ocean defines and structures this entire ecosystem, considered to be one of the largest marine biomes (Arístegui *et al.*, 2009). The mesopelagic zone (200-1000 m), where often the thermocline is located, shows a great variability in water masses and its associated physical parameters. This region is considered to be crucial in organic matter remineralization, showing marked peaks or deficits of oxygen and inorganic nutrients (Nagata *et al.*, 2010). Below, the bathypelagic zone (1000-4000 m) represents a much less variable environment. The physical conditions of this zone, in particular the low temperature (-1 to 3ºC), high pressure (10-50 MPa) and saturated oxygen concentrations, are globally quite stable suggesting a seemingly homogeneous habitat. Nevertheless, even in this region, it is possible to detect spatial gradients both for abiotic and biotic parameters caused by the different origins and properties of the bathypelagic water masses and by the inherent variability in the concentration and composition of organic constituents (Nagata *et al*., 2010). These gradients are expected to also influence the biological realm.

Given the absence of photosynthesis, microbial food webs in the dark ocean are sustained by imported organic matter from upper layers and prokaryotic production, including chemosynthetic reactions using reduced inorganic compounds such as ammonia or carbon monoxide (Dick *et al*., 2013). These reactions have an important effect on global carbon sequestration in the oceans (Jiao *et al*., 2011). In microbial food webs, heterotrophic protists (HP) are considered to be the first consumers of prokaryotic production. Whereas the importance of HP as grazers in surface waters is well established (e.g. Gasol *et al*., 2009), less is known about the magnitude of this function in deep waters. Some authors have proposed that protistan grazers play a minor role in controlling deep prokaryotic production (Nagata *et al*., 2010; Morgan-Smith *et al*., 2010; Boras *et al*., 2010), whereas others claim a significant grazing pressure on prokaryotes both in mesopelagic and bathypelagic layers (Fukuda *et al*., 2007; Arístegui *et al*., 2009; Cho *et al*., 2000). A first step towards solving this issue would be a quantification of the abundance and biomass of deep HP on a global scale.

Data on HP abundance in deep waters are available from various studies in separate marine regions: one Mediterranean site sampled at different times (Tanaka and Rassoulzadegan, 2002), four North Pacific stations (Yamaguchi *et al*., 2004), 6 Subarctic Pacific stations (Fukuda *et al*., 2007), 14 Pacific stations (Sohrin *et al*., 2010), 17 North Atlantic stations (Morgan-Smith *et al*., 2011) and 33 Equatorial Atlantic stations (Morgan-Smith *et al*., 2013). In general, these studies used epifluorescence microscopy to quantify HP. Microscopy may provide useful morphological

information (cell size, nucleus shape, presence of flagella) in the standard DAPI counts (Porter and Feig, 1980), or might allow to identificate specific taxonomic groups in FISH counts (Pernthaler *et al.*, 2002; Massana *et al.*, 2006), but is time-consuming and therefore limits the number of samples processed. Since flow cytometry (FC) counting is extremely useful for prokaryotes (Gasol and del Giorgio, 2000) and picophytoplankton (Dusenberry *et al.*, 1994), it would seem the right procedure for enumerating HPs as the method optimization was presented years ago (Zubkov *et al.*, 2006; Zubkov *et al.*, 2007) and refined recently (Christaki *et al.*, 2011). However, this approach has not yet been applied routinely to large-scale oceanographic surveys.

The aim of this paper is to report the abundance and biomass of heterotrophic protists in the dark deep ocean at a global scale, using both microscopy and flow cytometry. A large sampling effort was made during the Malaspina 2010 expedition, a circumnavigation cruise that sampled water masses down to 4000 m in the Atlantic, Pacific and Indian Oceans (Figure 1). Since concomitantly to the HP abundance we obtained abiotic parameters (temperature, oxygen, conductivity) and biotic parameters (viral abundance, prokaryote abundance and biomass), we could explore the relationship between HP and their environment.

Materials and methods

Sampling

The Malaspina 2010 Expedition on board the R/V BIO_Hesperides departed on December 2010 and finished on July 2011 and sampled a total of 147 stations around the world's main oceans. In this paper we present data from 116 stations (Figure 1). The cruise started and ended in the southern Iberian Peninsula and crossed the Atlantic, Indian and Pacific oceans. Mesopelagic and bathypelagic samples from at least five depths (between 200 and 4000 m) were collected with Niskin bottles attached to a rosette, which also had a Seabird 0911Plus CTD probe that measured temperature, salinity and oxygen along the vertical profiles. Seawater samples were prefiltered through a 200 µm mesh and then processed to estimate the abundance of HP by three different techniques: microscope counts by DAPI (4', 6- diamidino-2-phenylindole) staining, microscope counts by TSA-FISH (Tyramide Signal Amplification-Fluorescence In Situ Hybridization) using a eukaryote probe, and flow-cytometry counts. Samples for prokaryote and viral abundance were also collected.

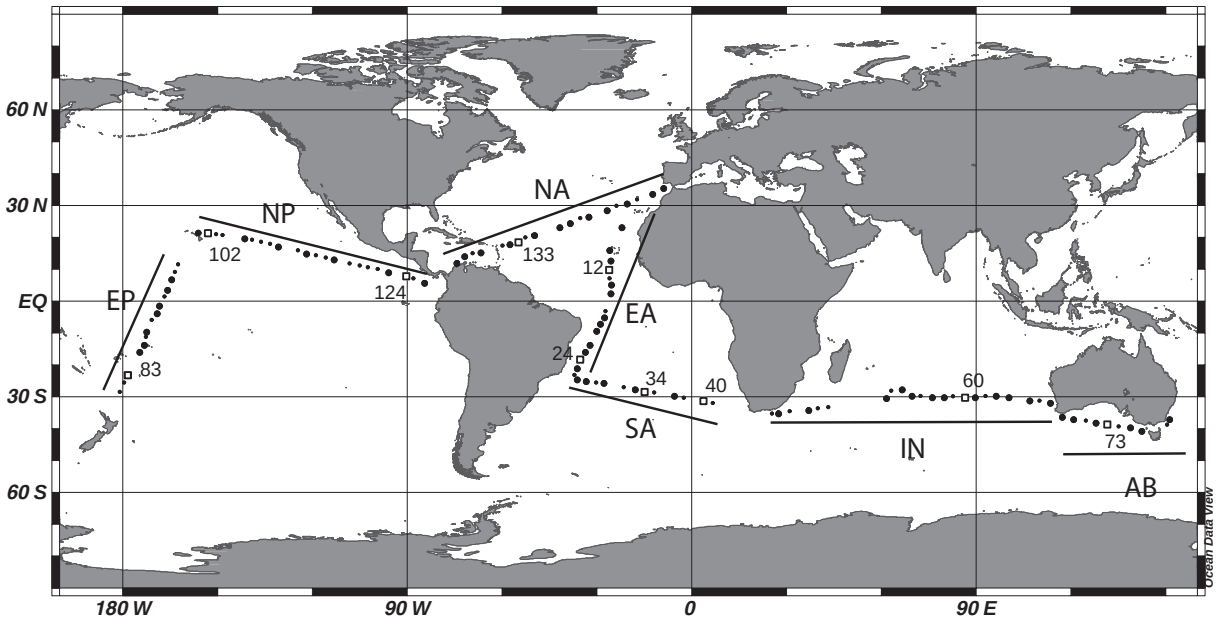


Figure 1 Map of the Malaspina 2010 cruise showing the 116 stations where the abundance of deep heterotrophic protists was measured. Small dots indicate stations where only the deepest sample was processed, large dots stations where the vertical meso- and bathypelagic profile was processed, and numbered squares stations used for microscopy. The cruise was divided in seven regions: Equatorial Atlantic (EA), South Atlantic (SA), Indian (IN), Great Australian Bight (AB), Equatorial Pacific (EP), North Pacific (NP), and North Atlantic (NA).

Epifluorescence microscopy counts by DAPI staining

Seawater samples were fixed with ice-cold 10% glutaraldehyde (1% final concentration), filtered on 0.6 μm pore-size polycarbonate black filters (25 mm) and stained with DAPI (0.5 mg ml^{-1}) (Porter and Feig, 1980). We filtered 27 mL of seawater for samples between 200 and 700 m and 180 mL for deeper samples. The filters were mounted on a slide with low-autofluorescence oil and stored at -20ºC in the dark until processed at the home institute within five months after the end of the cruise. Heterotrophic protists were counted with an epifluorescence microscope (Olympus BX61) at 1000x magnification by UV-excitation inspecting a transect of at least 20 mm (equivalent to 200 fields). Detected cells were inspected in blue light to confirm the lack of chlorophyll autofluorescence. At least 15 protist cells were counted per sample (average of 38 in all samples).

Epifluorescence microscopy counts by TSA-FISH

Samples for TSA-FISH were fixed with formaldehyde (1.85% final concentration) and filtered on 0.6 μm pore-size polycarbonate filters (25 mm). We filtered 95 mL of seawater for samples between 200 and 700 m and 475 mL for deeper samples. The filters were stored at -20ºC in the dark until processed at the home institute within five months after the end of the cruise. They were first embedded in 1% (w/v) low-gelling-point agarose to minimize cell loss. The hybridization was carried out by covering filter pieces with 20 μl of hybridization buffer (40% deionized formamide, 0.9 M NaCl, 20 mM Tris-HCl pH 8, 0.01% sodium dodecyl sulfate [SDS], and 20 mg ml^{-1} block-

ing reagent [Roche Diagnostic Boehringer]) containing 2 μl of HRP-labeled probe (stock at 50 ng μl^{-1}) and incubating at 35°C overnight. We used the oligonucleotide probe EUK502 (Lim *et al.*, 1999) that targets all eukaryotes. After two successive washing steps of 10 min at 37°C in a washing buffer (37 mM NaCl, 5 mM EDTA, 0.01% SDS, 20 mM Tris-HCl pH 8), the filters were equilibrated in PBS for 15 min at room temperature. Tyramide Signal Amplification was done for 30 to 60 min at room temperature in the dark in a solution containing 1x PBS, 2 M NaCl, 1 mg ml^{-1} blocking reagent, 100 mg ml^{-1} dextran sulfate, 0.0015% H_2O_2 and 4 µg ml^{-1} Alexa 488-labeled tyramide. The filters were then placed in PBS buffer twice for 10 min, rinsed with distilled water and air-dried. The cells were counterstained with DAPI (5 µg ml^{-1}) and the filter pieces were mounted with antifading mix (77% glycerol, 15% VECTASHIELD, 8% PBS 20x). Enumeration was done under blue light excitation using the same routine as above. We counted a minimum of 15 protist cells per sample (62 cells on average).

Pictures of heterotrophic protists visualized by TSA-FISH were taken with an Olympus DP72 camera connected to the microscope. Cell dimensions (in µm) were measured on the images with the Image Pro Plus software analyzer (Media Cybernetic Inc., Bethesda, MD, USA). Cell biovolumes (V, in μm^3) were calculated assuming prolate spheroid shapes (Hillebrand *et al.*, 1999) following the formula:

$$V = \pi/6 * d^2 * h$$

where h is the largest cell dimension and d is the largest cross section of h. We then used the equation of Menden-Deuer and Lessard (2000) to convert cell biovolume to cell biomass:

$$\text{Cell biomass (pg C cell}^{-1}) = 0.216 * V^{0.939}$$

Within each sample, average cell biomass times cell abundance counted by TSA-FISH resulted in the total biomass of the HP assemblage.

Flow cytometry counts

For protists, 4.8 mL of seawater were fixed with 25% glutaraldehyde EM grade (1% final concentration), deep-frozen in liquid nitrogen and stored at -80ºC until analyzed in the laboratory within seven months after the end of the cruise. Samples were processed with a FACSCalibur flow cytometer (BD-Biosciences) with a blue laser emitting at 488 nm using the settings explained by Christaki *et al.* (2011) adapted from the Zubkov *et al.* (2007) protocol. Each sample was stained for at least 10 min in the dark with DMSO-diluted SYBRGreen I (Molecular Probes, Invitrogen)

at a final concentration of 1:10000. The flow rate was established at about 250 mL min^{-1}, with data acquisition for 5-8 min depending on cell abundance. Samples showing more than 1200 events s^{-1} were diluted. The flow cytometer output was analyzed using CellQuest software (Becton Dickinson), initially visualized as a cloud of points in a window showing side scatter (SSC) versus green fluorescence (FL1), which contained all cells stained by SYBR Green I. From this plot, target cells were identified after excluding the noise, autofluorescent particles and heterotrophic prokaryotes, using different displays of the optical properties of the detected particles, as explained in Christaki *et al.* (2011).

For heterotrophic prokaryotes, 1.2 mL of seawater were fixed with a paraformaldehyde-glutaraldehyde mix (1% and 0.05% final concentrations, respectively) and stored as before . Samples were stained with SYBRGreen I, at a final concentration of 1:10.000, for 15 min in the dark at room temperature. The flow rate ranged between 35 mL min^{-1} (low) for samples above 1000 m depth, and 150 mL min^{-1} (high) for deeper samples. Acquisition time ranged from 30 to 260 seconds depending on cell concentration in each sample. Data was collected in a FL1 versus SSC plot and analyzed as detailed in Gasol and del Giorgio (2000). Polyscience latex beads (1μm) were always used as internal standards.

For viruses, 1.2 mL of seawater were fixed with glutaraldehyde (0.5% final concentration) and stored as before. Samples were stained with SYBRGreen I, and run at a medium flow speed after being diluted with TE buffer (1X Tris-EDTA) such that the event rate was between 100 and 800 viruses s^{-1} (Marie *et al.,* 1999). The data obtained for FL1 and SSC were collected and analyzed to select only the high DNA-content viruses (Large Viruses) from the total pool of viral particles (Brussaard *et al.,* 2004).

Cell biovolume of prokaryotes was estimated using the calibration obtained by Calvo-Díaz and Morán (2006) for oceanic samples, which relates relative side scatter (population SSC divided by beads SSC) to cell size. We used the same beads as in that study. Cell biovolume was converted to cell biomass with the equation of Gundersen *et al*. (2002):

$$\text{Cell biomass (fg C cell}^{-1}) = 108.8 * V^{0.898}$$

Results

Optimizing counts of heterotrophic protists by flow cytometry (FC)

We selected ten stations well distributed along the Malaspina cruise (numbered in Figure 1) to compare counts of heterotrophic protists by FC with those obtained by the time-consuming but presumably more accurate epifluorescence microscopy. The standard counting approach based on DAPI staining has the advantage that allows discriminating between nucleus and cytoplasm and often displays the presence of flagella, making the identification more accurate. On the other hand, TSA-FISH specifically targets protists (those cells having eukaryotic ribosomes), and large bacteria are not confounded. Therefore, it was chosen as a second standard to test and improve FC counts. Both microscopic counts provided very similar results (Figure 2a), with a linear slope of 1.02 ± 0.07, not significantly different than 1 ($p<0.0001$; n= 48; y=7.49) and a R^2 of 0.83.

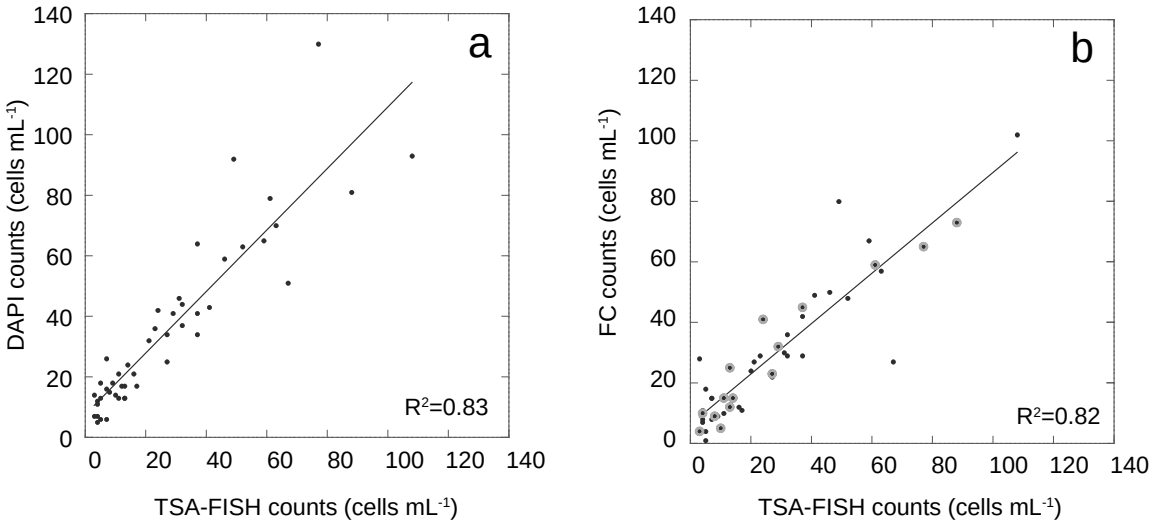


Figure 2 Methodological comparison of deep ocean heterotrophic protists counts. DAPI counts versus Flow Cytometry (FC) counts (**a**) and versus TSA-FISH counts (**b**) in samples from ten selected vertical profiles (shown as numbered stations in Figure 1). Samples in panel **b** used to position the FC window are encircled by a light grey area.

In FC counts, the accurate estimation of HP cells depends on how they are discriminated from heterotrophic prokaryotes in the cytograms, since both cell types are similarly labelled and share the same fluorescent properties (and differ by size). For different depth ranges (200-450, 451-700, 701-1400, 1401-4000) in three stations (40, 73 and 124), we identified the cytogram gate that displayed the best agreement between FC and TSA-FISH counts (linear slope of 0.81 ± 0.09, $p<0.0001$; R^2 of 0.91; n= 15; y=6.26; light grey dots in Figure 2b). This gate positioning was then applied to the samples from the other vertical profiles for which we had TSA-FISH data, and we obtained a very strong relationship between both counting methods in the ten stations (linear slope of 0.83 ± 0.07, $p<0.0001$; R^2 of 0.82; n= 48; y=6.26; Figure 2b). These FC settings were subse-

quently applied to the vertical profiles of the other 55 stations (large dots in Figure 1) and to the deepest sample of the remaining stations (small dots in Figure 1).

Altogether, we estimated the abundance of deep heterotrophic protists in 71 vertical profiles combining the information obtained by microscopy and flow cytometry (10 profiles by the three methods, 55 profiles by FC, and 6 profiles by TSA-FISH [in stations with inaccurate FC counts]) and in the deepest sample of 45 additional stations. In total, we estimated the HP abundance in 476 individual samples.

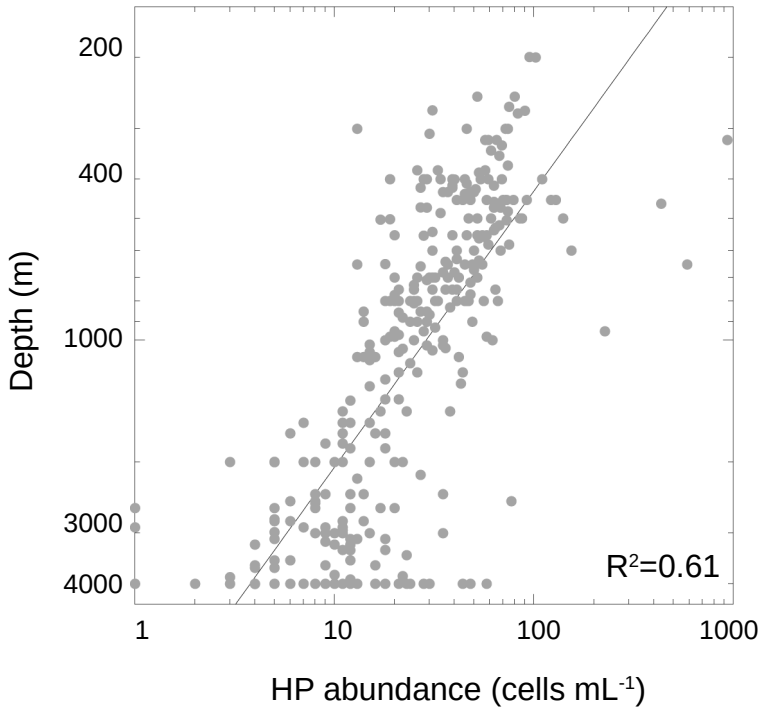


Figure 3 Abundance of heterotrophic protists versus depth in a log-log plot including all counts from this global study.

Main factors structuring HP abundance

We explored the possibility of predicting HP abundance with a multiple regression model using several abiotic parameters, such as depth, temperature, oxygen and salinity and one biotic variable, prokaryote abundance. We did not use Large Viruses abundance in this step due to the low number of vertical profiles processed (20 out of 71). After the first explorative analysis we maintained only the three parameters that showed significance ($p<0.05$): depth, oxygen and prokaryote abundance. Repeating the analysis with these variables only, they had a very strong effect on HP abundance, with a significance of $p<0.0001$ for depth and prokaryotic abundance and $p<0.001$ for oxygen. The entire model explained 66% of the variability. Looking at the beta coefficient of each variable, which represented their relative strength, showed that depth had the highest weight (beta=0.61), followed by prokaryotic abundance (0.28) and oxygen concentration (0.08). Next we will analyze the effect of the two main factors: depth and prokaryotic abundance.

Table 1 A global view of microbial components (protists, prokaryotes and large viruses) in the deep ocean.

	Heterotrophic Protists			Prokaryotes		Viruses
Depth m	*Abundance cells mL^{-1}*	*Biovolume µm^3 cell^{-1}*	*Biomass pg C mL^{-1}*	*Abundance 10^5 cells mL^{-1}*	*Biomass pg C mL^{-1}*	*Abundance 10^5 particles mL^{-1}*
200-450	72 ± 19	25 ± 5	280 ± 46	2.15 ± 0.26	837 ± 152	9.79 ± 1.43
451-700	70 ± 10	26 ± 5	150 ± 23	1.44 ± 0.09	661 ± 160	7.24 ± 0.91
701-1400	32 ± 3	32 ± 6	112 ± 28	0.98 ± 0.07	534 ± 106	3.91 ± 0.54
1401-4000	11 ± 1	39 ± 8	50 ± 14	0.56 ± 0.08	309 ± 59	1.47 ± 0.20

The table shows the average values and standard errors for abundance, cell biovolume and community biomass in four different depth layers. Values of abundance are referred to 71 vertical profiles (20 profiles in case of viruses), whereas values of biovolume and biomass derive only from 6 vertical profiles.

Heterotrophic protist abundance versus depth

Globally the abundance of HP decreased with depth with a log-log abundance versus depth slope of -0.68 ± 0.04 and a R^2 of 0.61 ($p<0.0001$; Figure 3). The average abundance of HP in the top layer of the mesopelagic region (200-450 m) was 72 cells mL^{-1} (±19) (Table 1). Cell abundances were very similar in the following depth layer, whereas further down into the bathypelagic region they halved to 32 cells mL^{-1} (±3) and reached an average of 11 cells mL^{-1} (±1) in the deepest layer (1401-4000 m).We did not detect significant differences in the abundance-depth slopes among the three oceans considered (slopes of -0.70 ± 0.04 in the Atlantic, -0.66 ± 0.05 in the Indian, and -0.66 ± 0.06 in the Pacific; data not shown), although this could vary at a regional scales. For instance the slope was -0.54 ± 0.06 in the North Atlantic region and -0.87 ± 0.08 in the South Atlantic samples. Similarly, the Equatorial Pacific region showed a slope of -0.77 ±0.11 and the North Pacific a slope of -0.80 ± 0.07.

A contour plot presenting HP abundance at all depths along the complete cruise track is shown in Figure 4. We divided the cruise into seven oceanic regions (see Figure 1) and, of those, the region with the highest HP abundance in the mesopelagic zone was the Equatorial Pacific (98 ± 38 cells mL^{-1} on average) (Table 2). An Analysis of Variance (ANOVA) test showed that the mesopelagic abundances in the Equatorial Pacific were significantly higher than those of the Atlantic and North Pacific Oceans ($p<0.05$). In bathypelagic waters, the highest HP abundance was attained also in the Equatorial Pacific region (20 ± 2 cells mL^{-1} on average). In this case, the averaged value was significantly higher than those at the North Pacific, Atlantic and Indian Oceans (ANOVA, $p<0.05$).

Heterotrophic protists counts were done at the deepest sample (ca 4,000 m) in the entire cruise (n=116). Abundances ranged between 1 and 58 cells ml^{-1}, and 75% of the counts were below 11

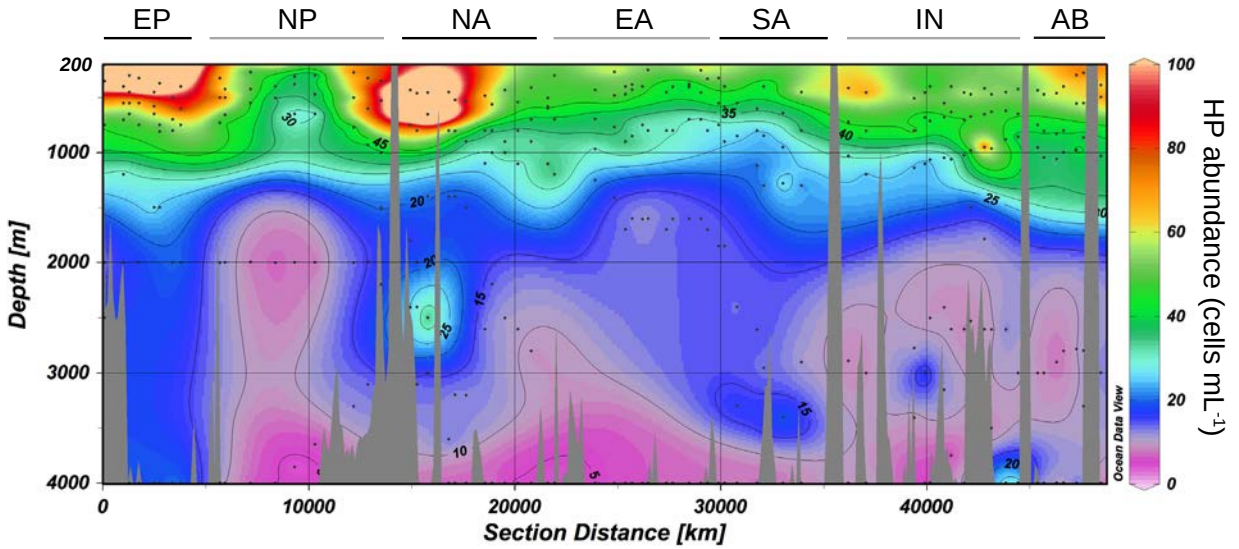


Figure 4 Abundance of heterotrophic protists with depth along the entire cruise obtained with ODV(Schiltzer, R., Ocean Data View, 2013). The track is separated in the oceanic regions indicated in Figure 1 (the departure and arrival harbor, Cadiz, appear in the middle of the plot only for graphical reasons). Small dots indicate sampling points.

cells ml^{-1} (Figure S1). As shown before, most samples from the Equatorial Pacific were above this value. Due to an overlapping between large prokaryotes and small-size protists, these samples were recounted by TSA-FISH microscopy, giving a robust support for the generally higher HP abundances in this oceanic region.

Using the surface area for Atlantic, Pacific and Indian Oceans, we calculated an approximate volume (in 10^7 km^3) for mesopelagic (6.59, 1.33, and 5.88, respectively) and bathypelagic layers (24.7, 49.7, and 22.1, respectively). Using these volume estimates and mean cell counts, we calculated the global number of cells for each ocean. Thus, for the mesopelagic layer, we found 7 x 10^{24} cells in the Pacific, 4 x 10^{24} in the Atlantic and 3 x 10^{24} in the Indian Oceans. For the bathypelagic layer, the Pacific and Indian displayed the same number than in the mesopelagic waters whereas for the Atlantic is 3 x 10^{24} .

Heterotrophic protist abundance versus prokaryote and viral abundance

The relationship between HP and prokaryotic abundance was carefully analyzed by comparing both estimates for all samples in the 71 vertical profiles (n=325). At this global level, the abundance of prokaryotes and that of HP was significantly correlated ($p<0.001$) with an R^2 of 0.50 and a log-log slope of 0.85 ± 0.05 (Figure 5a). However, this pattern varied in each particular oceanic region, with slopes ranging from 0.77 ± 0.10 in the South Atlantic to 1.28 ± 0.13 in the North Atlantic (Table 3). Again, the Equatorial Pacific was unusual, since the relationship between prokaryote and HP abundances in that ocean was not significant ($p= 0.08$).

The ratio between prokaryotes and HP abundance in the global dark ocean was 4251 (±237) (Table

Table 2 Microbial abundances of HP and prokaryotes (and the ratio between both estimates) in the seven oceanic regions shown in Figure 1.

Region	*Stations*		*HP abundance* cells mL^{-1}	*Prokaryotic abundance* 10^5 cells mL^{-1}	*Ratio Prok:HP*
EA	1-26	Total	29 ± 3	0.91 ± 0.09	3956 ± 414
		Meso	43 ± 3	1.33 ± 0.10	3217 ± 518
		Bathy	12 ± 2	0.40 ± 0.05	4866 ± 864
SA	27-41	Total	25 ± 3	0.46 ± 0.06	1950 ± 151
		Meso	38 ± 4	0.71 ± 0.09	2064 ± 247
		Bathy	14 ± 2	0.24 ± 0.03	1848 ± 185
IN	45-68	Total	32 ± 4	1.09 ± 0.10	4680 ± 558
		Meso	52 ± 7	1.54 ± 0.10	3391 ± 219
		Bathy	13 ± 2	0.67 ± 0.12	5930 ± 1039
AB	69-78	Total	34 ± 4	1.45 ± 0.16	6135 ± 1081
		Meso	53 ± 4	2.21 ± 0.15	4305 ± 184
		Bathy	13 ± 3	0.64 ± 0.09	8073 ± 2149
EP	81-98	Total	69 ± 23	1.33 ± 0.17	4109 ± 807
		Meso	101 ± 38	1.39 ± 0.18	2762 ± 484
		Bathy	20 ± 2	1.24 ± 0.31	6263 ± 1854
NP	101-126	Total	30 ± 4	1.79 ± 0.34	6097 ± 662
		Meso	43 ± 6	2.31 ± 0.45	5486 ± 927
		Bathy	9 ± 1	1.00 ± 0.49	6844 ± 937
NA	127-146	Total	43 ± 13	0.58 ± 0.06	3055 ± 406
		Meso	80 ± 28	0.94 ± 0.09	2346 ± 228
		Bathy	15 ± 3	0.30 ± 0.02	3603 ± 688
Global	1-146	Total	34 ± 3	0.99 ± 0.05	4251 ± 237
		Meso	54 ± 5	1.44 ± 0.06	3371 ± 175
		Bathy	14 ± 1	0.51 ± 0.04	5177 ± 439

The table shows the average values and standard errors in the total deep region or in the mesopelagic and bathypelagic layers.

2). The ratio was lower in the mesopelagic region, 3364 (±174), than in the bathypelagic region, 5195 (±441). There were significant differences between oceans (Table 2), with minimal ratios in the South Atlantic (1848 ± 185) and maximal ratios in the Great Australian Bight (8073 ± 2149).

Total viral abundances were obtained in only 20 vertical profiles, which were then used to investigate their relationship with HP abundance. Several viral fractions were measured, and for the purpose of this paper we considered only viral particles with large genome size (Large viruses, LV), which comprise generally viruses infecting protists. LV and HP abundance were well correlated (Figure 5b), with a statistically significant slope in the log-log plot ($p<0.001$). LV abundance alone explained almost 30% of the variance (R^2 of 0.28) with a slope of 0.33 ± 0.05. Moreover, this correlation was also significant considering the three oceanic regions separately: Atlantic (slope=0.33 ± 0.09, $p<0.001$, $R^2=0.20$), Indian (slope=0.63 ± 0.13, $p<0.001$, $R^2=0.44$), Pacific (slope=0.37± 0.07, $p<0.001$, $R^2=0.56$). We performed a multiple regression analysis to identify the relative im-

Table 3: Slopes of the log-log relationships between the abundances of prokaryotes and HP, with additional statistics, for each oceanic region defined in Figure 1

Region	*Slope*	*p*	R^2
EA	1.05 ± 0.08	0.0001	0.75
SA	0.77 ± 0.10	0.0001	0.64
IN	1.16 ± 0.10	0.0001	0.69
AB	1.14 ± 0.12	0.0001	0.74
EP	0.33 ± 0.18	0.0757	0.08
NP	0.95 ± 0.12	0.0001	0.61
NA	1.28 ± 0.14	0.0001	0.62

portance of LV in this dataset of 20 vertical profiles. The entire model explained 53% of the variability. Taking a look on single variables, the relation with prokaryotic abundance was still highly significant ($p<0.0001$) with a beta coefficient of 0.58 whereas for LV this value was less than half (0.21) and had lower significance ($p=0.034$).

Cell size and biomass estimations

In seven selected stations (24, 34, 60, 73, 83, 102, and 133) we measured the size of individual cells by processing TSA-FISH microscopic images in all samples of the vertical profile. No clear differences were seen when comparing vertical profiles, and here we present the data together. We calculated the average cell biovolume of HP in the same depth layers defined before (Table 1). In the upper layer, the mean cell biovolume was 25 μm^3, which was slightly lower than that of the deeper layer (39 μm^3), although the differences were no significant ($p=0.09$; t student). Within the

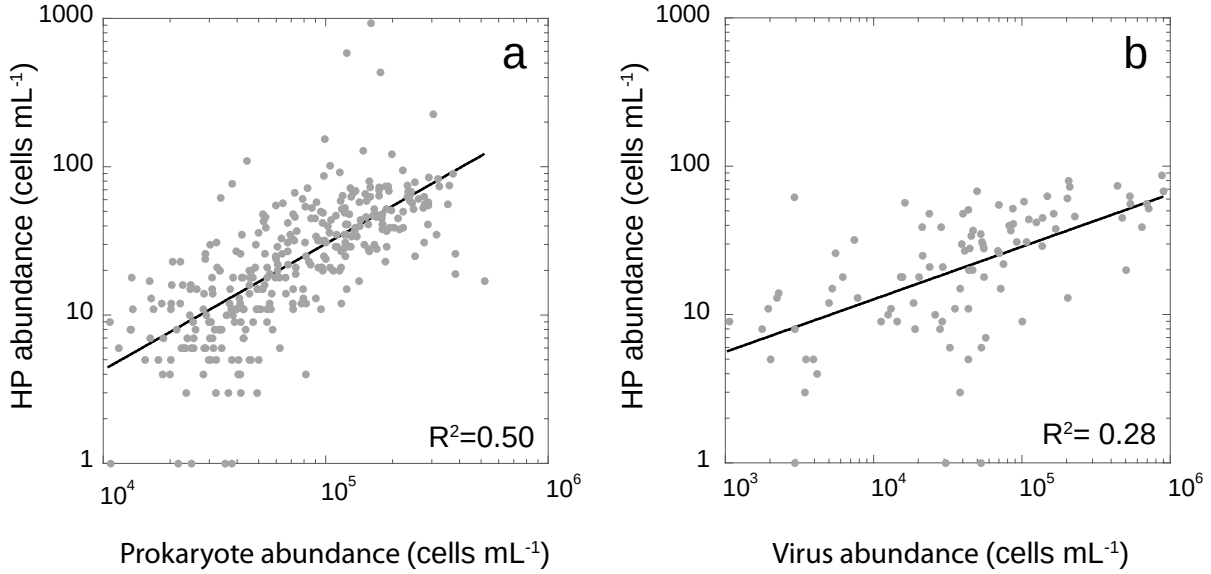


Figure 5 Abundance of heterotrophic protists versus prokaryote abundance (**a**) and large viruses abundance (**b**), in samples deriving from 71 and 20 vertical profiles, respectively.

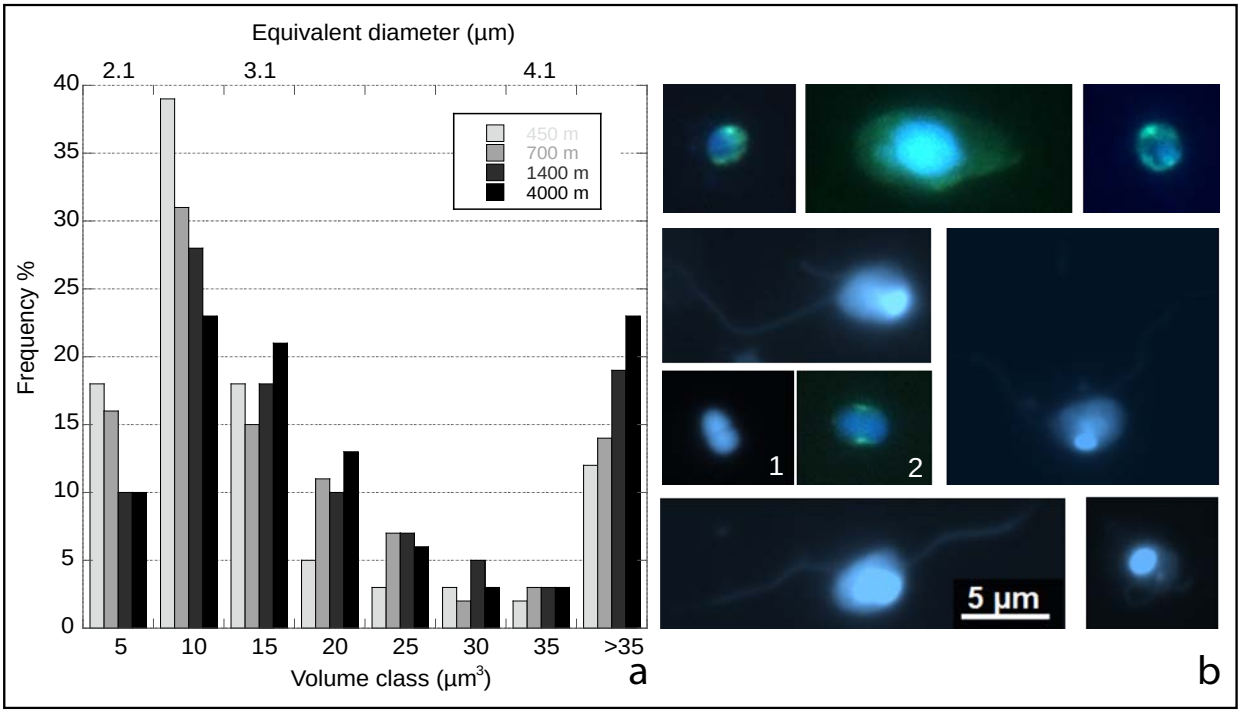


Figure 6 (**a**) Cell biovolume spectra of HP cells in different depth layers. The equivalence of cell biovolume to equivalent spherical diameter is also indicated. (**b**) Some micrographs of bathypelagic HP cells, showing different cell shapes and the presence of flagella. The blue signal corresponds to the DAPI-stained nucleus, and the green signal to the TSA-FISH stained cytoplasm. Split morphotypes are shown in pictures 1 and 2.

cell size spectra, the most frequent classes were those between 10 and 15 µm^3 (Figure 6a). The number of very small cells (equivalent diameter <3 µm) decreased with depth. Thus, 75% of cells in the 200-450 m layer were below this size threshold, whereas this value was 62%, 57%, 54% in the consecutive depth layers. Some images of protist cells used to measure cell dimension are shown in Figure 6b.

We then used the mean cell biovolume to calculate cell biomass and, together with cell abundances, the community HP biomass for the seven vertical profiles. HP biomass ranged from 4 to 486 pg C mL^{-1} (Figure S2a). Grouping the vertical profiles in the same four depth layers as before, we found an average value of 280 pg C mL^{-1} in the upper 200-450 m layer, and a subsequent reduction of HP biomass in the three following layers: 150, 112, and 50 pg C mL^{-1} (Table 1). Three bathypelagic samples showed deviating high values: station 102 in North Pacific at 2000 m (146 pg C mL^{-1}), station 73 in the Great Australian Bight at 2800 m (175 pg C mL^{-1}) and station 60 in the Indian Ocean at 4000 m (90 pg C mL^{-1}). In the last two cases, higher biomass values were due to larger cell sizes, and not to higher abundances.

The biomass of prokaryotes in the same seven vertical profiles also decreased with depth, but the decrease was less pronounced than that of HP biomass (Figure S2b). The slopes of the log-log plot were -0.53 for HP biomass and -0.75 for prokaryotes, and they were significantly different ($p<0.0001$, ANCOVA). Consequently, the log-log plot of prokaryotic versus HP biomass using all

samples revealed a no significant relation (p=0.09). However, this relationship becomes significant when removing station 73 (with anomalous high biomass) from the analysis (n=29, slope of 0.84, p=0.01, R^2=0.22). The global ratio between eukaryotic and prokaryotic biomass was 0.30 (± 5), being 0.39 for the mesopelagic and 0.21 for the bathypelagic.

Discussion

We provide a general picture of the abundance of heterotrophic protists in mesopelagic and bathypelagic waters of the world's main oceans. Compared with the research done on prokaryotes, only a handful of studies have enumerated deep HP (Pomeroy and Johannes, 1968; Sorokin *et al.*, 1985; Tanaka and Rassoulzadegan, 2002; Yamaguchi *et al.*, 2004; Fukuda *et al.*, 2007; Sohrin *et al.*, 2010; Morgan-Smith *et al.*, 2011; Morgan-Smith *et al.*, 2013) likely due to the difficulty of sampling very deep in the ocean and to the time-consuming enumeration techniques required. It is important to highlight that the magnitude of our sampling effort (116 stations) and the geographical coverage in our study (see Figure 1) are larger than all the previous studies together, therefore allowing for a more refined picture of the distribution of these deep microbes.

We aimed at using flow cytometry to estimate the abundance of HP (Christaki *et al.*, 2011), a routine that had not yet been used in large-scale oeanographic surveys. In parallel, we used microscopy in selected samples to test the accuracy of flow cytometry, verify FC counts, and exclude unrealistic values. Deep HP visualized in DAPI-stained preparations included several cell shapes and the presence of flagella, but sometimes their identification was doubtful. This led us to use the TSA-FISH technique with a probe targeting the whole eukaryotic community, to complement the general DAPI-staining. The relation between the two methods (epifluorescence and FC) was very good (Figure 2). Coupling techniques combining the speed of automatic enumeration with the accuracy of direct observations is strongly recommended in case of a large number of samples as typically derived from oceanographic cruises.

At a global level, the abundance of heterotrophic protists decreased from 72 cells mL^{-1} in the upper mesopelagic layer to 11 cells mL^{-1} at the lower bathypelagic layer, with a log-log slope of -0.68 ± 0.04. This decrease was very similar to that found in a previous review (Arístegui *et al.*, 2009), where HP decreased with depth with a slope of -0.66 considering Atlantic and Pacific samples. Despite the global decrease trend, the distribution of HP cells was not equal at the same depth range over the analyzed transect (Figure 4). These ocean basin variations in HP abundance depended mostly on prokaryote abundance and on oxygen concentration. The importance of large viruses

(LV), as suggested before (Wommack *et al*., 1999; Steward *et al*., 2000), was analyzed separately with a subset of 20 stations, and showed that the LV abundance explained less than half the variability of prokaryotes. In general, the HP abundance observed in the bathypelagic layer (ca 1-15 cells mL^{-1}) were in the same range than previous reports (Fukuda *et al.*, 2007; Tanaka and Rassoulzadegan, 2002; Sohrin *et al.*, 2010; Boras *et al.*, 2010; Morgan-Smith *et al.*, 2013). However, along the entire expedition we found several exceptional points, particularly in the South Pacific.

During epifluorescence inspections it was possible to identify most of the cell shapes defined by Morgan-Smith *et al.* (2011, 2013). Although counting given shapes was not the aim of this study, we noticed that the "split morphotype" (with no clear taxonomic assignation), which was the most abundant morphotype in that study, was almost ubiquitous in Malaspina bathypelagic samples (Figure 6b, see pictures 1 and 2). Many of the microscopically observed cells showed flagella (Figure 6b), suggesting they were active bacterial grazers (Jürgens and Massana 2008). With respect the mean size, deep flagellates tended to be slightly larger than surface ones. Indeed, 54% of the bathypelagic protists had a biovolume between 5 and 15 μm^3 (Figure 6a) corresponding to spherical equivalent diameters of 2 to 3 μm, while this value in surface waters was about 76% (Jürgens and Massana, 2008). Although not significant, the mean cell biovolume tended to increase with water layer depth. In particular, the number of cells larger than 35 μm^3 (>4 µm in diameter) represented 12% at 250 m and 22% at 4,000 m. This is in contrast with Fukuda *et al*. (2007), who found a decrease in the contribution of larger cells with depth in the subarctic Pacific. The absence of deformed of exploded cells during the microscopic counts led us to exclude the effects of volume enlargement due to decompression.

The estimations of HP community biomass were done in one vertical profile per oceanic region, where cell size was measured, and allowed to infer a general trend. As expected, at a global level, HP biomass decreased clearly with depth, from 280 pg C mL^{-1} at the upper mesopelagic layer to 49 pg C mL^{-1} at the lower bathypelagic layer. The average biomass for the bathypelagic realm was one order of magnitude larger than the values estimated by Fukuda *et al*. (2007) and Sohrin *et al*. (2010) but similar to other reports (Tanaka and Rassoulzadegan, 2002; Yamaguchi *et al*., 2004). The biomass ratio between HP and prokaryotes was 0.21 ± 0.05 in the global bathypelagic realm, whereas it was 0.39 ± 0.08 in mesopelagic realm. This is reflected by a faster decrease of HP biomass than prokaryote biomass (slopes of -0.53 and -0,75, respectively). The excess of prokaryote biomass (as compared to HP biomass) lets open the question about the importance of the grazing pressure in the deeper bathypelagic ocean.

The impact of grazing on prokaryotes in the deep ocean is still a matter of debate (Fukuda *et al*.,

2007; Arístegui *et al.*, 2010; Nagata *et al.*, 2010; Boras *et al.*, 2010), and interesting clues can derive from analyzing the ratio in the abundance of prokaryotes and HP cells (PROK:HP ratio). Considering the bathypelagic region globally there were 5195 (± 441) prokaryotic cells for each protist. This is three times the ratio found in an epipelagic reference dataset, 1760 ± 162 (data were collected from the following papers: Kirchman *et al.*, 1989; Cho *et al.*, 2000; Tanaka and Rassoulzadegan, 2002; Yamaguchi *et al.*, 2002 and 2004; Tanaka *et al.*, 2005), indicating less protists for a given prokaryote cell abundance in deep waters than at surface. A possible explanation could be a lower cell-specific prokaryote production in deep waters as compared with surface, which would then support fewer protist cells in the deep ocean (Arístegui *et al.*, 2009). In addition, the protistan grazing rate is generally correlated with temperature (Vaqué *et al.*, 1994) so lower grazing rates in the deep and cold ocean could again result in higher PROK:HP ratios. Another putative reason for this finding would be that the prokaryote abundance in the deep ocean is below the numerical threshold of grazing (Andersen and Fenchel, 1985), so protists spend too much energy (via respiration) in the search for prey and as a result prokaryotes are inefficiently grazed resulting in higher PROK:HP ratios. Despite low prokaryote abundances, HP cells could still be sustained given the micropatch distribution theory (Simon *et al.*, 2002; Baltar *et al.*, 2009) that suggests that most of the interactions between HP and prokaryotes take place in aggregates where prey density is high enough to sustain HP growth.

Therefore, on a global scale the PROK:HP ratio is clearly higher in the deep ocean than at surface. Interestingly, though, this ratio displays a substantial variability at local scale. For instance, the Atlantic community is characterized by a low ratio with no significant difference between mesopelagic and bathypelagic regions, while the bathypelagic layer of the Great Australian Bight exhibits the highest ratios (8073 ± 2149). In order to seek for an explanation for this variability, we analyzed the presence of fungal signal in the deep ocean, derived from a parallel study of the diversity of deep microeukaryotes done by pyrosequencing (Pernice *et al.*, in preparation). In general, sites with high ratios, such as the Pacific, showed a larger contribution of fungi. The relationship between the PROK:HP ratio and the percentage of fungal sequences (Figure 7) was very significant ($n=20$, $p=0.0003$, $R^2=50$). Fungal species are known to be osmotrophs, consuming organic matter directly from the environment and not by ingesting particulate material such as prokaryotes by phagocytosis (Richards *et al.*, 2011). The presence of fungi within deep HP assemblages would mean that prokaryotes are not the only carbon source and several microeukaryotic clades could be instead osmotrophs. The fact that some of the HP counted are osmotrophs instead of prokaryote grazers could result in a certain relaxation of the grazing pressure on prokaryotes, thus deriving in higher PROK:HP ratios. In particular, perhaps the higher vertical flux of organic matter in the

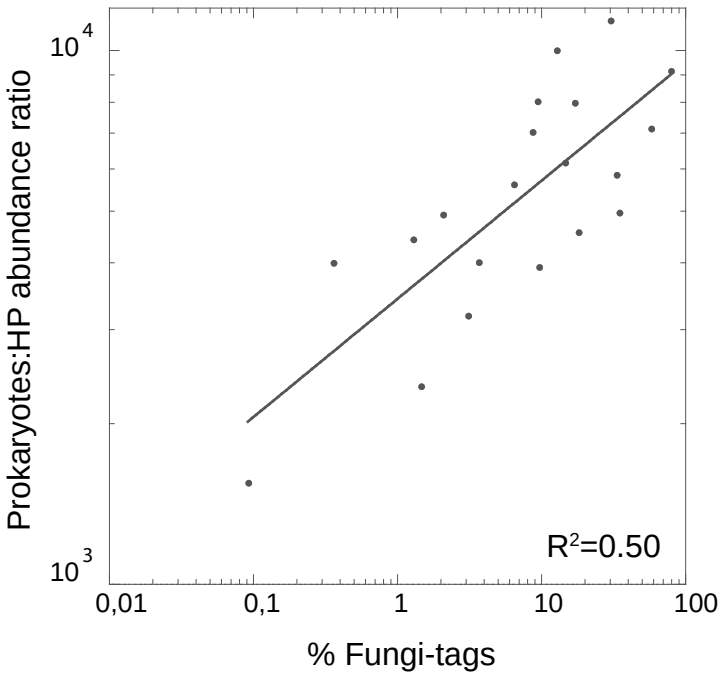


Figure 7 Relationship of the ratio in the abundances of prokaryotes and HP cells with respect the percentage of Fungi sequences in the corresponding samples. The later values derives from a parallel study on deep protist diversity (Pernice et al, in preparation).

South Pacific Ocean (Hansell *et al.,* 1997) could uncouple the relationship between prokaryotes and HP, since some of them could grow directly on sedimenting organic matter. The good relationship shown in Figure 7 is supporting the hypothesis that osmotrophy (as estimated by the relative abundance of fungal sequences) is explaining higher PROK:HP ratios.

This study confirms and extends previous results on the HP distribution in the deep ocean, and provides a more refined global view. Our wide sampling coverage showed that HP were ubiquitous, with minimal abundances around 10 cells mL^{-1}, and that their biomass averaged approximately 20% of prokaryote biomass in the global bathypelagic realm. The maintenance of this microeukaryotic biomass likely requires active grazing on prokaryotes or the presence of osmotrophic processes. Our work suggests that we should consider HP important players in the dark ocean and highlight the importance of studying the dynamics and diversity of this microbial food web component.

Acknowledgements

This study was supported by the Consolider-Ingenio Malaspina 2010 financed by the former Ministry of Science and Innovation (MICINN). We thank the scientists that sampled for heterotrophic protists in the different legs of the cruise: Roy Mackenzie, Laura Alonso-Sáez, Eva Sintes, Mireia Mestre, Paqui García and Txetxu Arrieta.

References

Andersen P, Fenchel T (1985). Bacterivory by microheterotrophic flagellates in seawater samples. *Limnol Oceanogr* **30:**198-202

Arístegui J, Duarte CM, Gasol JM, Herndl GJ (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnol Oceanogr* **54:** 1501-1529.

Baltar F, Arístegui J, Sintes E, Van Aken HM, Gasol JM, Herndl GJ (2009). Evidence of prokaryotic metabolism on suspended particulate organic matter in the dark waters of the subtropical North Atlantic. *Limnol Oceanogr* **54:** 182-193.

Boras JA, Sala MM, Baltar F, Arístegui J, Duarte CM, Vaqué D (2010). Effect of viruses and protists on bacteria in eddies of the Canary Current region (subtropical northeast Atlantic). *Limnol Oceanog* **55:** 885-898.

Brussaard CPD (2004). Optimization of procedures for counting viruses by flow cytometry. *Appl Environ Microbiol* **70:** 1506-1513.

Calvo-Díaz A, Morán XAG, Suárez LÁ (2008). Seasonality of picophytoplankton chlorophyll a and biomass in the central Cantabrian Sea, southern Bay of Biscay. *J Mar Syst* **72:** 271-281.

Cho BC, Na SC, Choi DH (2000). Active ingestion of fluorescently labelled bacteria by mesopelagic heterotrophic nanoflagellates in the East Sea, Korea. *Mar Ecol Prog Ser* **206:** 23-32.

Christaki U, Courties C, Massana R, Catalá P, Lebaron P, Gasol JM *et al* (2011). Optimized routine flow cytometric enumeration of heterotrophic flagellates using SYBR Green I. *Limnol Oceanogr: Methods* **9:** 329-339.

Dick GJ, Anantharaman K, Baker BJ, Li M, Reed DC, Sheik CS (2013). The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front Microbiol* **4:** 124.

Dusenberry JA, Frankel SL (1994). Increasing the sensitivity of a FACScan flow cytometer to study oceanic picoplankton. *Limnol Oceanogr* **39:** 206-209.

Fukuda H, Sohrin R, Nagata T, Koike I (2007). Size distribution and biomass of nanoflagellates in meso- and bathypelagic layers of the subarctic Pacific. *Aquat Microb Ecol* **46:** 203-207.

Gasol JM, del Giorgio PA (2000). Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. *Sci Mar* **64:** 197-224.

Gundersen K, Heldal M, Norland S, Purdie DA, Knap AH (2002). Elemental C, N, and P cell content of individual bacteria bollected at the Bermuda Atlantic Time-Series Study (BATS) Site. *Limnol Oceanogr* **47:** 1525-1530.

Hillebrand H, Dürselen C-D, Kirschtel D, Pollingher U, Zohary T (1999). Biovolume calculation for pelagic and benthis miceoalgae. *J Phycol* **35:** 403-424.

Hansell D A, Carlson C A, Bates N R, Poisson A (1997). Horizontal and vertical removal of organic carbon in the equatorial Pacific Ocean: a mass balance assessment. *Deep Sea Res* **44:** 2115:2130

Jiao N, Zheng Q (2011). The microbial carbon pump: from genes to ecosystems. *Appl Environ Microbiol* **77:** 7439-7444.

Jürgens K, Massana R (2008). Protistan grazing on marine bacterioplankton. In: Kirchman DL (ed). *Microbial Ecology of the Oceans*, Second edition. John Wiley & Sons, Inc., New York, USA, pp 383-441.

Kirchman D, K'nees E, Hodson R (1985). Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. *Appl Environ Microbiol* **49:** 599-607.

Kirchman DL, Keil RG, Wheeler PA (1989). The effect of amino acids on ammonium utilization and regeneration by heterotrophic bacteria in the subarctic Pacific. *Deep-Sea Res* **36:** 1763-1776.

Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D (1999). Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl Environ Microbiol* **65:** 45-52.

Menden-Deuer S, Lessard EJ (2000). Carbon to volume relationship for dinoflagellates, diatoms and other protist plankton. *Limnol Oceanogr* **45:** 569-579

Morgan-Smith D, Herndl GJ, van Aken HM, Bochdansky AB (2011). Abundance of eukaryotic microbes in the deep subtropical North Atlantic. *Aquat Microb Ecol* **65:** 103-115.

Morgan-Smith D, Clouse MA, Herndl GJ, Bochdansky AB (2013). Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. *Deep-Sea Res Part I: Oceanographic Research Papers* **78:** 58-69.

Nagata T, Tamburini C, Arístegui J, Baltar F, Bochdansky AB, Fonda-Umani S *et al* (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochem-

istry, and genomics. *Deep-Sea Res Part II: Topical Studies in Oceanography* **57:** 1519-1536.

Pernthaler J, Glöckner F, Schönhuber W, Amann R (2001). Fluorescence in situ hybridization (FISH) with rRNA-targeted oligonucleotide probes. *Method Microbiol* **30:** 207-226.

Pomeroy LR, Johannes RE (1968). Respiration of ultraplankton in the upper 500 meters of the ocean. *Deep-Sea Res* **15:** 381-391.

Porter K, Feig Y (1980). The use of DAPI for identifying and counting aquatic microflora. *Limnol Oceanogr***:** 943-948.

Richards TA, Jones MDM, Leonard G, Bass D (2012). Marine Fungi: Their Ecology and Molecular Diversity. *Ann Rev Mar Sci* **4:** 495-522.

Simon M, Grossart H-P, Schweitzer B, Ploug H (2002). Microbial ecology of organic aggregates in aquatic ecosystems. *Aquat Microb Ecol* **28:** 175-211.

Smith DC, Azam F (1992). A simple, economical method for measuring bacterial protein synthesis rates in seawater using 3H-leucine. *Mar Microb Food Webs* **6:** 107-114.

Sohrin R, Imazawa M, Fukuda H, Suzuki Y (2010). Full-depth profiles of prokaryotes, heterotrophic nanoflagellates, and ciliates along a transect from the equatorial to the subarctic central Pacific Ocean. *Deep-Sea Res Part II: Topical Studies in Oceanography* **57:** 1537-1550.

Sorokin YI (1985). Phosphorus metabolism in planktonic communities of the Eastern Tropical Pacific. . *Mar Ecol Prog Ser* **27:** 87-97.

Steward GF, Montiel JL, Azam F (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* **45:** 1697-1709.

Tanaka T, Rassoulzadegan F (2002). Full-depth profile (0–2000 m) of bacteria, heterotrophic nanoflagellates and ciliates in the NW Mediterranean Sea: Vertical partitioning of microbial trophic structures. *Deep-Sea Res Part II: Topical Studies in Oceanography* **49:** 2093-2107.

Tanaka T, Rassoulzadegan F, Thingstad TF (2005). Analyzing the trophic link between the mesopelagic microbial loop and zooplankton from observed depth profiles of bacteria and protozoa. *Biogeosciences* **2:** 9-13.

Vaqué D, Gasol JM, Marrasé C (1994). Grazing rates on bacteria: The significance of methodology and ecological factors. *Mar Ecol Prog Ser* **109:** 263-274.

Wommack EK, Ravel J, Hill RT, Chun J, Colwell RR (1999). Population dynamics of Chesapeake Bay virioplankton: total-community analysis by pulse-field gel electrophoresis. *Appl Environ Microbiol* **65:** 231-240.

Yamaguchi A, Watanabe Y, Ishida H, Harimoto T, Furusawa K, Suzuki S *et al* (2002). Structure and size distribution of plankton communities down to the greater depths in the western North Pacific Ocean. *Deep-Sea Res Part II: Topical Studies in Oceanography* **49:** 5513-5529.

Yamaguchi A, Watanabe Y, Ishida H, Harimoto T, Furusawa K, Suzuki S *et al* (2004). Latitudinal differences in the planktonic biomass and community structure down to the greater depths in the western north Pacific. *J Oceanogr* **60:** 773-787.

Zubkov MV, Burkill PH (2006). Syringe pumped high speed flow cytometry of oceanic phytoplankton. *Cytometry Part A* **69A:** 1010-1019.

Zubkov MV, Burkill PH, Topping JN (2007). Flow cytometric enumeration of DNA-stained oceanic planktonic protists. *J Plankton Res* **29:** 79-86.

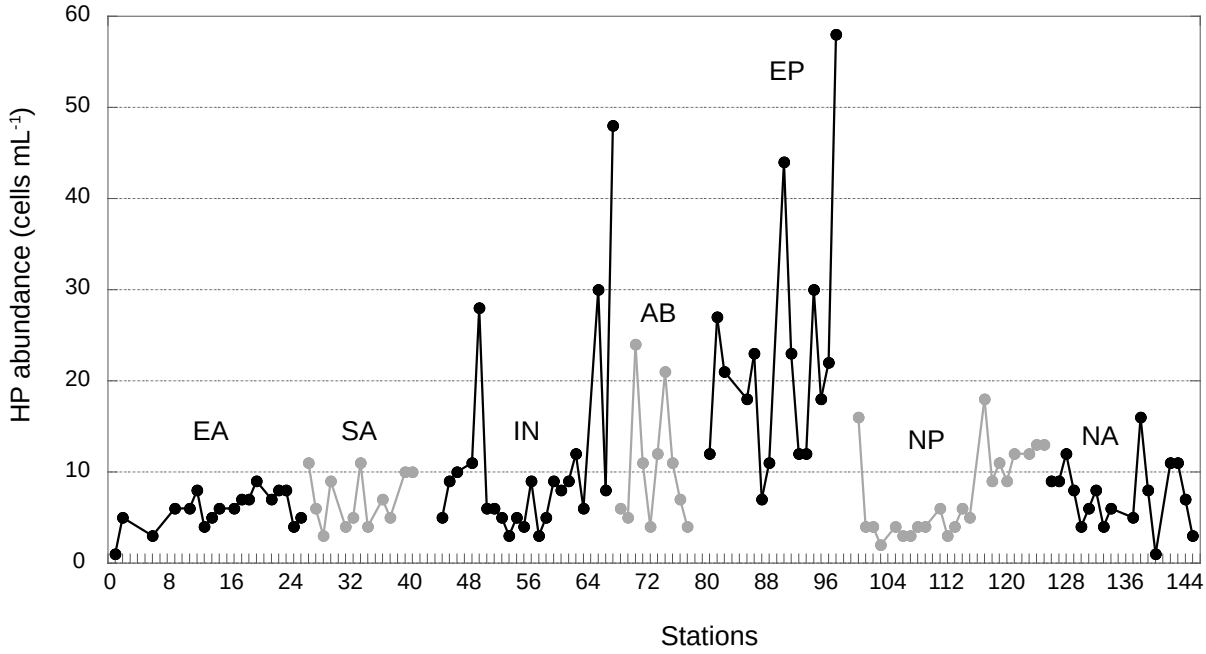


Figure S1 Abundance of heterotrophic protists during the entire Malaspina 2010 cruise in the deepest sample analyzed (generally at 4,000 m). The seven oceanic regions depicted in Figure 1 are shown in the graph.

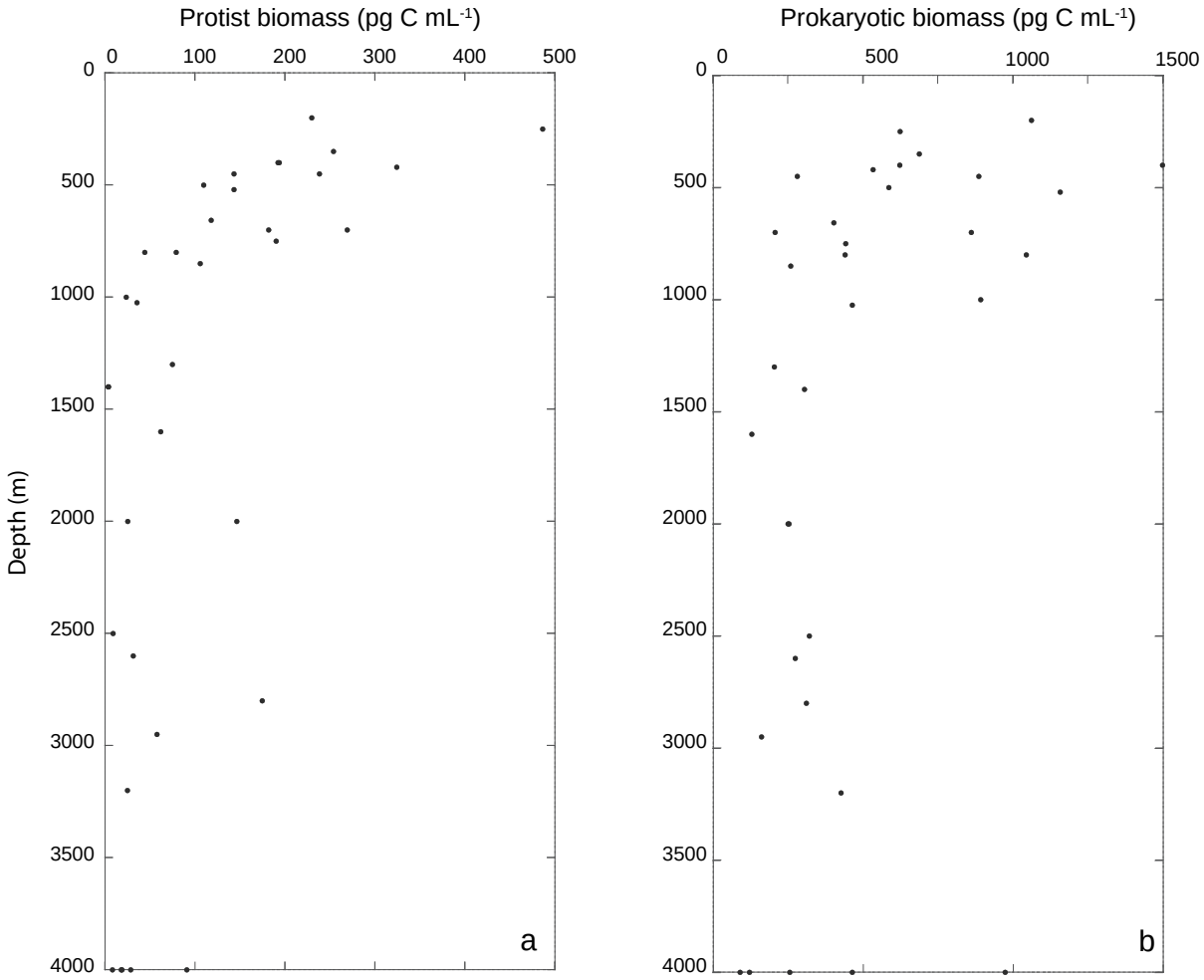


Figure S2 Community biomass of HP (**a**) and prokaryotes (**b**) in samples from seven vertical profiles

Chapter 4

Diversity of marine microeukaryotes in the global deep ocean

Pernice MC, Rodríguez C, Logares R, Perera J, Acinas S, Gasol JM, Massana R. Diversity of marine microeukaryotes in the global deep ocean. In preparation.

Abstract

The aim of this work is to study the diversity of bathypelagic microeukaryotes. Seawater samples (3000 to 4000 m depth) came from 27 stations of the Malaspina-2010 global expedition that covered the Atlantic, Pacific and Indian Oceans. Pyrosequencing was used to obtain more than half a million tags from the 18S rDNA V4 region that after several curation steps (removal of low-quality sequences, short tags, OTUs occurring in a single sample, and chimeras) were clustered into 2482 OTUs at 97% similarity. The relative pyrotag abundance of the 20 most abundant OTUs matched well with the results of a parallel metagenomic analysis of 18S rDNA genes, suggesting that the tag-approach was little affected by PCR biases. There was a weak trend of genetic similarity among geographically close stations and among samples from the same water-mass. In addition, the ratio in cell abundance between prokaryotes and microeukaryotes had a significant relation with taxonomic composition. Despite 42 OTUs were found in all samples, there was not a typical global community. Instead, there were four main phylogenetic groups (Collodaria, Chrysophytes, Basidiomycota and MALV-II) mixed in different proportions. The amount of phylogenetic novelty was concentrated in three hotspots, one in each ocean, and accounted for 6% of pyrotags globally. Rarefaction curves suggested that there were species still waiting to be discovered. Our study is the essential first step for a more detailed investigation of the deep ocean microbiota and suggests idiosyncratic microeukaryotic assemblages in distinct regions.

Introduction

The bathypelagic region of the deep ocean, defined as the water column between 1000 and 4000 m, comprises a huge biome in terms of extension but it is as unknown as the moon surface. Generally, the range of variability of its abiotic parameters is narrow, apparently defining a very stable environment (Angel 1993). Thus, pressure for a given depth is constant, as well as temperature (range of -1 to 3 ºC), salinity (34.3 to 35.1) and dissolved oxygen concentration (2.4 to 5.7 mg L^{-1}). Nevertheless, several other parameters, such as the concentration of inorganic nutrients or particulate and dissolved organic matter, known to play an important role and building the niche structure for microbial life, fluctuate at broad regional scales depending on the flux of organic components from the surface and the occasional presence of hydrothermal vents (Nagata *et al.* 2010). Considering the importance of the deep ocean in biogeochemical cycles, particularly its role in organic matter remineralization and carbon reservoir, and the contribution of microorganisms in these processes, it seems critical to characterize each element of the deep microbial assemblages.

Despite the deep ocean can be regarded as an extreme environment, characterized by a low energy income, this ecosystem holds many and varied life forms, mostly microbial, which trophically interact in the well studied microbial food web (Azam *et al.* 1983, Fuhrman *et al.* 1992, Michaels and Silver 1988, Vaqué *et al.* 1994). Besides unpigmented prokaryotes, the second most apparent component of the system are heterotrophic microeukaryotes or heterotrophic protists (HP), generally considered as bacterial grazers. Globally, the averaged HP abundance in bathypelagic waters is 14 ± 1 cells mL^{-1}, representing a biomass of 50 ± 14 pg C mL^{-1} (Pernice *et al.*, submitted).

A few papers have analyzed the diversity of bathypelagic microeukaryotes by using clone libraries of 18S rDNA genes and Sanger sequencing, an approach that is inherently limited in the amount of sequences generated. Some studies analyzed the microeukaryotes in the water column, always on a regional scale, both by using universal eukaryotic primers (Lopez-Garcia *et al.* 2001, Stoeck *et al.* 2003, Countway *et al.* 2007, Not *et al.* 2007) or group-specific primers (Bass *et al.* 2007, Lara *et al.* 2009). Other diversity studies were done in sediments (Edgcomb *et al.* 2011a, Salani *et al.* 2012) or in hydrothermal vents (Edgcomb *et al.* 2002, Sauvadet *et al.* 2010). The bathypelagic HP diversity has also been studied by FISH staining and automatic microscopic inspections (Morgan-Smith *et al.* 2011, Morgan-Smith *et al.* 2013). Although FISH provides useful quantitative information of given taxa, it remains restricted to the taxa targeted by existing probes, and therefore never targets all known groups. Moreover, probe design often requires the existence of previous sequencing information.

High throughput sequencing now allows a much more exhaustive assessment of microbial diversity. Tag sequencing using the 454-pyrosequencing technique (Margulies *et al.* 2005) provides orders of magnitude more sequences than the Sanger method, and has been used targeting the 18S rDNA gene in studies of marine (Edgcomb *et al.* 2011b, Logares *et al.* 2012, Kilias *et al.* 2013) and freshwater (Charvet *et al.* 2012) surface microeukaryotes. This approach includes a PCR step, known to be prone to a series of biases like DNA polymerase errors or primer selectivity, which may affect final amplicon ratios (Acinas *et al.* 2005). Interestingly, metagenomics now allows a PCR-free approach to microbial diversity, based on extracting 18S rDNA sequences (miTags when using Illumina) from the pool of environmental sequences (Logares *et al.* 2013). The combination of tag-sequencing with metagenome analysis on a set of samples from the circumnavigation cruise Malaspina-2010, which had the principal aim of studying the deep ocean at a global scale, would allow to shed light onto the global diversity of deep heterotrophic microeukaryotes.

Here we extracted environmental DNA, and pyrosequenced 18S rDNA genes, of marine deep microeukaryotes (0.8-20 µm size-fraction) from 27 stations located in the Atlantic, Indian, and Pacific Oceans. After several automatic and manual quality controls on the initial sequencing pool, we generated an OTUs table formed by 359,163 pyrotags clustered into 2482 OTUs at 97% similarity. The OTU table was first used to detect the differences between communities and to identify the environmental parameters driving these differences. Then it was used for the description of taxonomic diversity with particular attention to the dominant phylogentic groups in the deep environment. These groups were then confirmed by a parallel metagenomic analysis performed with the same samples. Our study is the first attempt to describe exhaustively the diversity of heterotrophic microeukaryotes in the bathypelagic ocean by using for the first time a high-throughput sequencing method on samples from a global geographic effort.

Materials and methods

Sampling

Sampling was done in 27 stations of the world oceans during the Malaspina 2010 circumnavigation (Figure 1), performed between December 2010 and July 2011 on board the R/V BIO Hesperides. Seawater samples were collected from bathypelagic depths with Niskin bottles attached to a rosette that also contained a Seabird 0911Plus CTD probe, which measured temperature, salinity and oxygen concentration. Forteen samples were collected at 4000 m, eleven at depths between 3000 and 4000 m, and two were shallower: 2400 m in station 62 and 2150 m at station 82. Seawa-

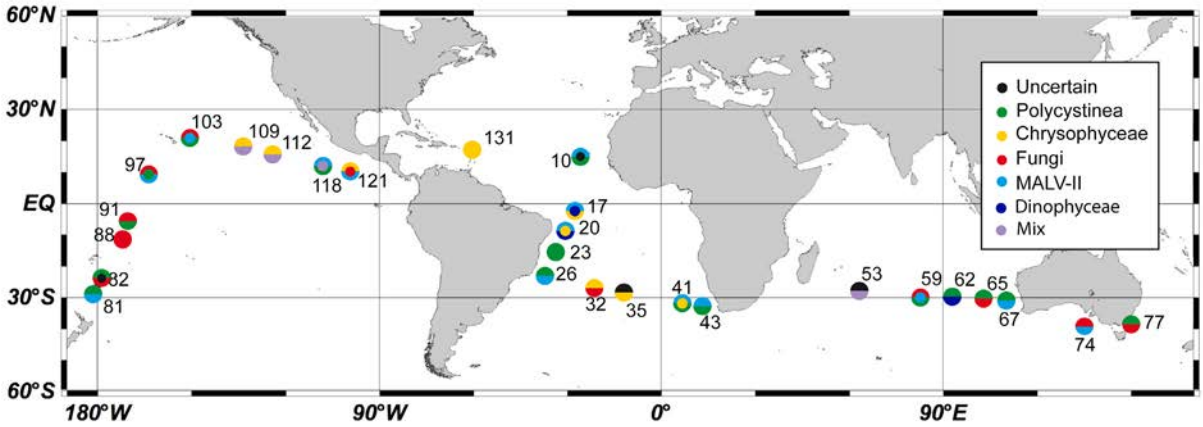


Figure 1. Map of stations sampled during the Malaspina 2010 expedition used for diversity analyses of bathypelagic protists. Each point is colored differently according to the dominant phylogenetic groups in the sample: one colour if a single group represents more than 75% of the pyrotags, two (or three) colours if the sum of the groups is more than 50%. Groups with more than 20% are always displayed; the category mix combines groups at lower abundance. The top half circle indicates the group at higher abundance, followed by the lower half circle and the inner circle.

ter was first prefiltered through a 200 μm mesh placed at the end of the hose and second through a 20 μm mesh in a funnel. Between 100 and 120 L of the 20 μm filtrate was then filtered with a peristaltic pump on 142 mm Millipore polycarbonate filters of 0.8 μm pore-size. Filters were flash-frozen in LN2 and stored at -80 until processed in the lab.

Along the Malaspina cruise, bathypelagic samples were a mixture of three principial water masses: North Atlantic deep water (NADW), Weddel Sea deep water (WSDW), and Circumpolar deep water (CDW). The proportion of these three water masses in each sample was inferred from measures of temperature, salinity and oxygen in the sampled seawater (J. Salgado, personal information). Samples were then clustered together in a dendrogram using these inferred proportions with Pvclust (Suzuki and Shimodaira 2006), to identify the water masses types sampled during the cruise. Six types were identified, NADW pure, NADW enriched, CDW pure, CDW enriched, CDW-WSDW, and WSDW enriched.

DNA extraction

Filters were cut into small pieces and soaked in 3 ml of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose). The protocol of DNA extraction was as follows. First, an enzymatic digestion started by incubating with Lysozyme (1 mg ml^{-1} final concentration) at 37ºC for 45 min while slightly shaken. Then, Proteinase K (0.2 mg ml^{-1} final concentration) and sodium dodecyl sulfate (1% final concentration) were added and filter pieces were incubated at 55ºC for 60 min while slightly shaken. The lysate underwent two steps of standard phenol-chloroform extraction to remove lipids and proteins. After the last centrifugation the aqueous phase was collected, concentrated in an Amicon ® Ultra unit (Millipore) and washed three times with 2 ml sterile deionized

water. After the third wash, between 100 and 250 µl of purified total genomic DNA extract was recovered and quantified using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies Inc, Wilmington, DE, USA). Typical yields were between 0.08 and 0.58 μg DNA, corresponding to ~1.8 ng per L of seawater.

Amplification of 18S rDNA genes and 454-sequencing

A two-step PCR was required to obtain enough DNA for pyrosequencing. We added 2 ng of genomic DNA to triplicate PCR tubes containing dNTPs (0.2 mM), the eukaryotic primers TAReukFWD1 (5'-CCAGCASCYGCGGTAATTCC -3') and TAReukREV3 (5'-ACTTTCGTTCTTGATYRA-3') at 0.5 mM (Stoeck *et al.* 2010) and the PCR buffer (1x) in a final volume of 20 µL. The typical amplicon size was about 380 bp. The initial PCR conditions consisted of an initial denaturation step of 98°C for 2 min and 30 s and then 10 cycles of 45 s at 98ºC, 35 s at 53ºC, and 35 s at 72ºC. Triplicate amplicons were pooled and purified using the QIAquick PCR Purification Kit to a final volume of 18 μL. Then, we did a second PCR step of 20 cycles with 2 μL of the previous PCR concentrate (45 s at 98ºC , 35 s at 48ºC, 30 s at 72ºC; with a final step of 10 min at 72°C) with newly added primers that were the same than before except that the forward primer had the 454 specific adaptor. Once proved with an agarose gel that this second PCR worked, we did 20 cycles in quintuplicate, each with 2 $_{\mu L \text{ of template}}$ and therefore using most of the previous PCR concentrate (10 μL). This increased DNA concentration without losing genetic diversity. Final PCR products were purified, eluted in 30 μl of sterile deionized water and the DNA was quantified with the Qubit 1.0 fluorometer (Invitrogen). About 200 ng of PCR product were sent for amplicon sequencing on a 454 GS FLX Titanium system (Lifesequencing S. L., Valencia, http://www.lifesequencing.com, Spain).

Processing 454 sequences (pyrotags) datasets

Raw 454 data was processed with QIIME (Caporaso *et al.* 2010) for demultiplexing and sequence quality control. Due to the version of the 454-sequencer, it was not possible to run DeNoiser or AmplicoNoise, so we decided to be very strict in the cleaning process and keep only the highest quality pyrotags. Selected sequences were from 150 bp to 600 bp in size, with no more than 2 mismatches in the primer, and without homopolymers longer than 8 bases. Then, the quality score in each position was averaged in running windows of 50 bp and pyrotags were truncated at the limit of the window having an average score lower than 25. Pyrotags resulting to be shorter than 150 bp were then removed. An OTUs table was constructed by clustering high-quality pyrotags from the complete dataset at a 97% similarity threshold. This OTUs table was manually curated by removing OTUs represented by short sequences (less than 250 bp) and OTUs that occurred

in only one sample, regardless their abundance (Figure 2). A representative sequence of each OTU (chosen as the most abundant) was then taxonomically assigned by using three reference databases: SILVA 108 (Quast *et al.* 2013), PR2 (Guillou *et al.* 2013), and MAS9013, an in-house database based in our previous work (Pernice *et al.* 2013). SILVA 108 was used to exclude 16S rDNA prokaryotic OTUs, whereas PR2 and MAS9013 were used to classify eukaryotic OTUs to established taxonomic groups, mostly at the class level. OTUs were assigned to a given group when its representative sequence had an e-value below 10^{-50} (equivalent to >90% similarity) with a reference sequence. Above this e-value, OTUs were classified as Uncertain. Chimera check was done with UCHIME (Edgar *et al.* 2011) using the MAS9013 database as reference with default parameters, and the results were carefully evaluated to avoid removing novel OTUs at this step. Thus, OTUs identified as chimera were kept if their representative sequence had an e-value of 0 against the GenBank database, the similarity between parents was above 90%, occurred in at least 14 stations or represented at least 100 pyrotags.

Statistical tools based on R environment were used to analyze this large amount of data. The comparative diversity analysis and the relation of diversity with environmental parameters were performed with the packages vegan (NMDS, Adonis test, Shannon index, star plot) and gmt (Geo-distance test) (Okasanen *et al.* 2013, Wessel *et al.* 2013). Considering the variability in the number of pyrotags among samples (from 6625 to 29,926) most comparative analyses were done on a subsampled set of the OTUs table, done with rrarefy, an additional tool of the R package vegan.

Analysis of 18S rDNA from metagenomes: miTags

Genomic DNA of the same size fraction (0.8-20 µm) from the same deep samples (except stations 74, 88, and 109) was used to obtain the metagenomes by Illumina sequencing (Bentley *et al.* 2008), done at the JGI under the project Deep Malaspinomics. Illumina reads containing 18S rDNA genes (miTags) were extracted from the entire pool of genes with a Hidden Markow Model software (HMMER v.3.1, Finn *et al.* 2011) using a default reference database (SILVA 108) for identification (Logares *et al.*, 2013). Only reads longer than 100 bp were kept (19,434 miTags). These were used as query against the dataset of 454 sequences, to compare its phylogenetic distribution. In a second and more refined step, we assessed the recovery of the 20 most abundant 454-OTUs in the miTag pool. MiTags retrieved by more than one OTU were assigned to the most similar one. Both analyses were done with BLAST constraining the similarity to 97% in a fragment of at least 100 bp. Comparisons were always done by averaging the relative contributions in the same set of 24 samples.

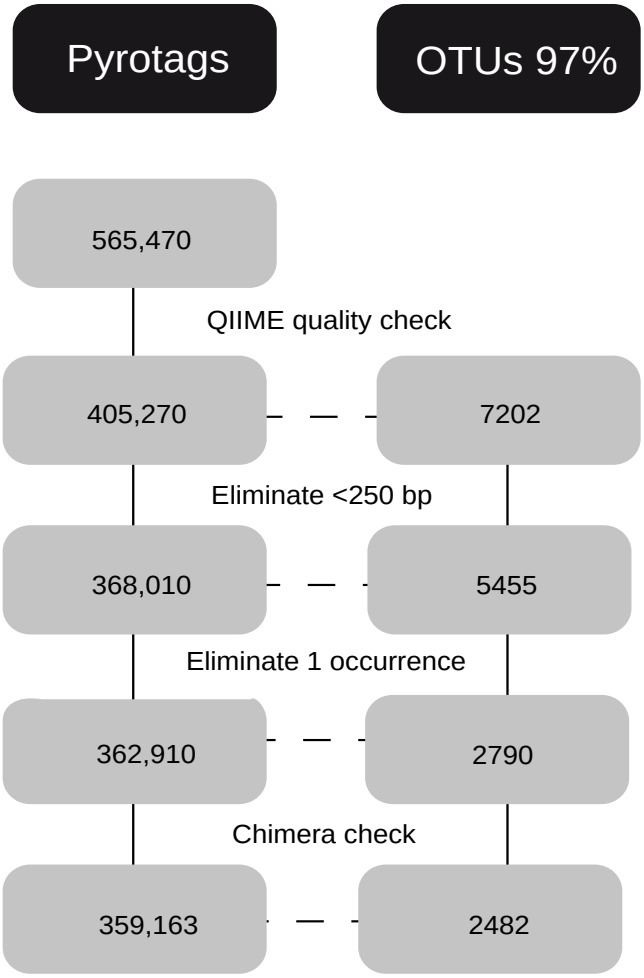


Figure 2. Overview of the cleaning steps in the dataset of 454 sequences (pyrotags), showing the number of pyrotags on the left and the corresponding number of OTUs clustered at 97% similarity on the right.

Results

In the frame of Malaspina-2010 cruise we collected deep samples, most of them at 3000 to 4000 m depth, from 27 stations in different regions of the world ocean (Figure 1). The diversity of deep microeukaryotes in these samples (size fraction of 0.8 to 20 µm) was then analyzed by pyrosequencing 18S rDNA genes. The output of the QIIME pipeline was an initial OTU table of 405,270 pyrotags clustered at 97% similarity in 7202 OTUs (Figure 2). This value was reduced to 362,910 pyrotags after eliminating short and OTUs occurring in only one sample. A posterior chimera check caused the loss of of about 10% of OTUs and 1% of pyrotags. The final OTU table included 2482 OTUs that represented 359,163 pyrotags. This OTU-table was then used to analyze first the community diversity in - and between- stations (alpha and beta diversity), and second the taxonomical affiliation of deep protists. The comparative community analysis included 25 samples (the less deep samples at stations 62 and 82 were excluded) while the complete set of 27 samples was considered for the taxonomical analysis.

Alpha diversity of deep protist assemblages

We performed rarefaction curves to check if the diversity saturated in any of the analyzed samples (Figure 3a). In general, the richness observed was between 400 and 1000 OTUs, and most samples did not show a sign of saturation (excepting perhaps samples from stations 23 and 131). As an estimate of alpha diversity in each sample we chose the Shannon index, and this ranged from 0.90 to 4.84 (Figure 3b).

These were calculated on an OTU table subsampled to have the same number of pyrotags (6625) per sample. A Mantel test indicated that this subsampling represented very well the original pool of sequences (R^2= 0.96, p<0.001, Figure S1). Considering oceanic regions separately, Shannon indices were similar: 3.62 (± 0.45) in the Atlantic, 3.56 (± 0.17) in the Indian and 3.69 (± 0.18) in the Pacific. At a more local scale, samples with the higher Shannon values (above 4.5) were in the Atlantic (stations 10, 17, 19 and 43), but this ocean also included the two samples with lowest diversity (stations 23 and 131). Distinct water masses also did not show distinctive Shannon indices (data not shown).

Beta diversity of protist assemblages

To shed light onto how similar was protist diversity among the 25 deepest samples we performed a non-metric multidimensional scaling (NMDS) analysis (Figure 4a), based on Bray-Curtis distances calculated from the subsampled OTU table of 6625 pyrotags per sample. Several NMDS plots were performed and the one displaying the minimal stress (0.17) was kept, as suggested before (Clarke et al., 1993). In an attempt to identify the physical parameters affecting the cluster-

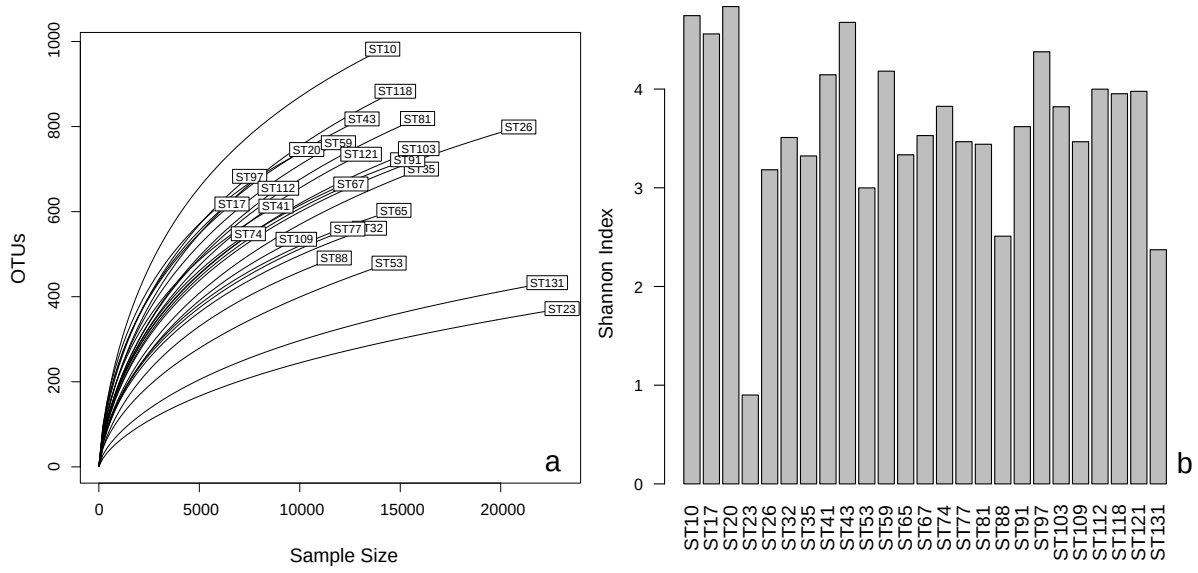


Figure 3. Alfa diversity features of deep protist assemblages as inferred by the analysis of pyrosequences clustered at 97% similarity **(a)** Rarefaction curves for each sample (**b**) Shannon indices calculated for each sample.

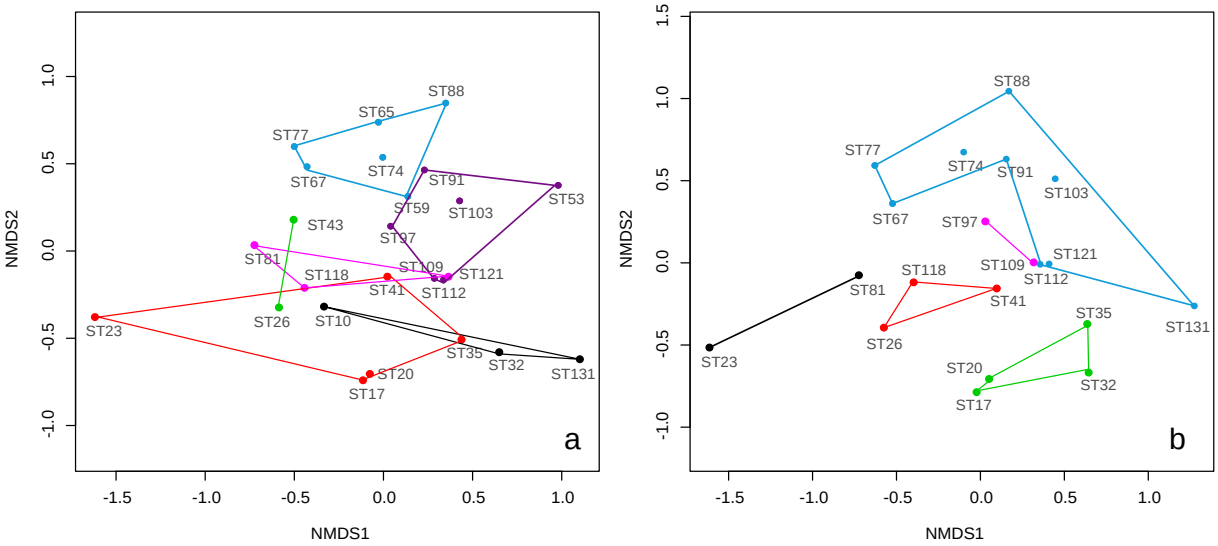


Figure 4. NMDS plots displaying the deep samples analyzed according to protist diversity similarities. Samples are then colored by several properties, in an attempt to interpret sample similarity. **(a)** NMDS plot with the complete set of the 25 deepest samples, grouped according to the water mass type: NADW pure (black), NADW enriched (red), CDW pure (pink), CDW enriched (violet), CDW-WSDW (light blue), WSDW enriched (green). **(b)** NMDS plot with the 20 stations for which we had the ratio in abundance of prokaryotes:protists. Stations were grouped by the values of this ratio: <2500 (black), 2501-4000 (red), 4001-4950 (green), 4951-10000 (light blue), >10000 (pink).

ing observed in the NMDS plot, we first colored the 25 stations and link them by boxes in base of their adscription to the defined water mass types (Figure 4a). Then, we performed an Adonis test with two continuous variables (prokaryotes abundance and bottom distance) plus two categorized variables (water mass and oceanic region) to identify the contribution of these factors in explaining sample organization (Table 1). The order of parameters matters so we followed a subordination criterion (prokaryotes abundance <water mass <bottom distance <ocean). Water mass explained 28% of the variability with high significance (p=0.007), while the other three variables were not significant (Table 1). Thus, the variance not explained by this set of variables was 72%.

In a second attempt, we did another analysis with the 20 samples for which we had the abundance ratio of prokaryotes and micreoeukaryotes, to use this as an additional explanatory factor of the variability observed. In the new NMDS we colored the samples in five categories of this ratio (Figure 4b), and apparently this explained much better the sample organization in the plot. Within the Adonis analysis (Table 1), this ratio explained 34% of the variability with high significance (P=0.0002), while water mass explained an additional 26% but with lower significance (p=0.03) and the ocean region 6% more. Taking into account this new variable, we could explain 66% of the variability.

The NMDS analysis showed a tendency of closer stations to be closer in the plot, and to better analyze the effect of geographic distances on community composition we performed a Mantel test to compare the distance matrix based on the OTU table against the matrix of geographic distances

Table 1. On the right, Adonis Test of station between 3000 and 4000 m, four parameters are tested two continuos (Prokaryotes abundance and Bottom distance) and two categorized (water mass and oceanic region), the water masses clustering is the same of the picture of NMDS, the oceanic regions are three (Atlantic, Indian and Pacific). The order of parameter matters and follows a dependence criterion. On the right, Adonis test for a subset of 20 stations where was available the ratio prokaryotes:heterotrophic protist abundance. The ratio was categorized in five groups (<2500, 3000-4000, 4001-4950, 4951-9500, >10000)

25 Stations			20 Stations		
Variable	R^2	p	Variable	R^2	p
Prok. abund.	0.06	0.058	Prok:HP	0.34	0.001
Water mass	0.28	0.007	Water mass	0.26	0.032
Bottom dist.	0.04	0.185	Bottom dist.	0.06	0.088
Ocean	0.05	0.194	Ocean	0.06	0.037

(Figure 5). The two matrices were related with a high significance (p=0.0001), although the Mantel test only explained 10% of the variability. In fact, geographically distant samples were almost always different, while closer samples were not always similar. We plotted the LOESS line (locally weighted scatter plot smoothing) to better evidence the relation of the genetic distance with the geographic distance at restricted spatial scales. The genetic distance increased rapidly among samples up to 600 km apart, but after this point the increase was weak.

Taxonomic identification of deep protists

Taxonomic identifications were performed at two taxonomic scales (supergroup and class-like group) and at two geographical scales (global and local). The global analysis of pyrotags at the supergroup level suggested that the deep ocean was dominated by Rhizaria (29%) and Alveolata (25%), followed by Fungi (14%) and Stramenopiles (14%) (Figure 6a). This was calculated by subsampling each sample to the same number of pyrotags (6625). Pyrotags included in the Uncertain category (sequences with similarity below 90% against reference databases) represented 6% of the total. The analysis of OTU_{97} was quite different (Figure 6b): Alveolata represented 52% of the OTUs, tripling the number of Rhizaria OTUs (17%), while Stramenopiles and Fungi had 4% and 3%, respectively. Interestingly, Excavata, with only 1% of pyrotags, accounted for 8% of the OTUs.

We identified 49 different taxonomic groups barely at class level, including MAST and MALV ribogroups and the uncertain category (Table 2). At a global level, Collodaria (Rhizaria, Radiolaria) was the group with the highest representation (78,394 pyrotags, on average 18.3% per sample), followed by MALV-II (Alveolata; 13.1%), Chrysophytes (Stramenopiles; 12.2%), Basidiomycota (Fungi; 10.9%), Dinoflagellates (Alveolata; 7.9%) and Uncertain (6.2%). It is interesting to note that four of the five most abundant groups belonged to different Supergroups.

To better display the differences in taxonomic composition among samples (local scale analysis),

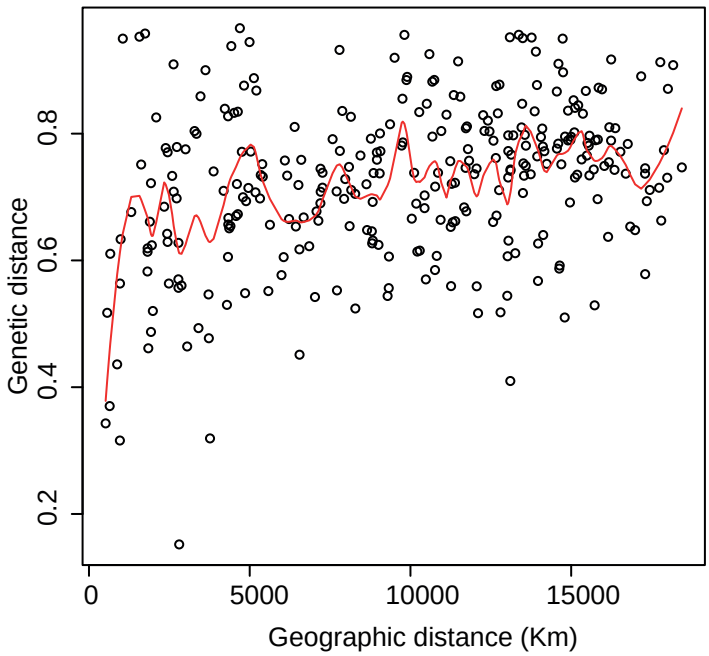


Figure 5. Mantel test relating Bray-Curtis genetic distances of the deep protists assemblages to geographic distance among samples. The line interpolating the values (LOESS line) is also shown.

we selected the 10 most abundant groups (including the Uncertain) and displayed their relative pyrotag contribution in samples grouped by water mass types (Figure 7). Fungi appeared to thrive in Pacific and Indian deep waters almost exclusively. Taken Ascomycota and Basidiomycota together, they represented on average 23% of pyrotags in CDW enriched samples and 28% in CDW-WSDW samples (with a maximal of 70% in sample 88), while they were less important in CDW pure samples (9%) and scarce in the pool of Atlantic samples (<5%). Polycystinea (Collodaria and Spumellaria together) were dominant in stations belonging to different water masses, being better represented in WSDW enriched (42% on average), CDW pure (32%), and in the mix of these two waters (32% in CDW-WSDW). Collodaria dominated in the first two water types, and Spumellaria in the third one. Chrysophytes preferred NADW pure (40% of pyrotags, with a peak

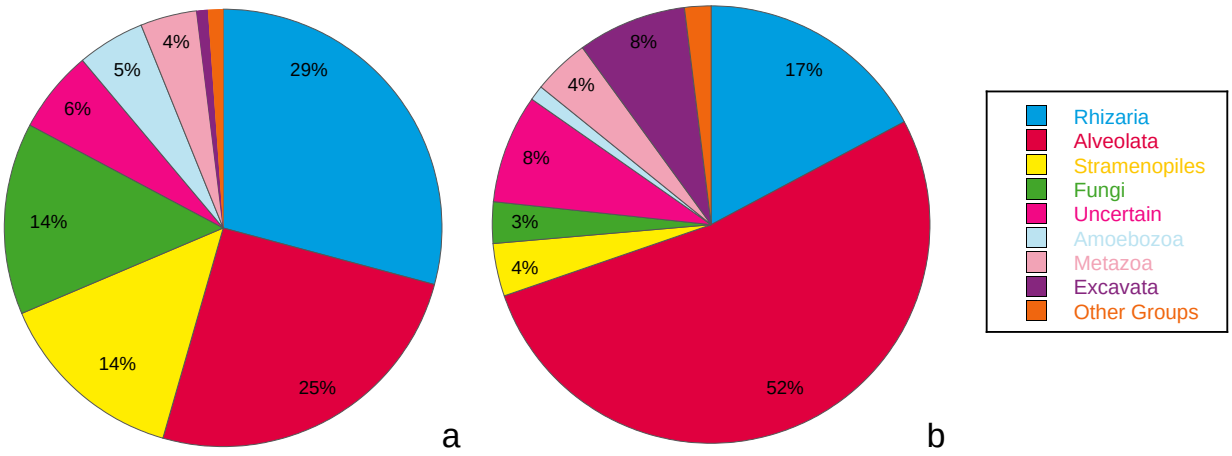


Figure 6. Overview of the diversity of deep protists at a supergroup taxonomic level. (**a**) Number of pyrotags per supergroup, averaging the relative abundance in each sample. (**b**) Number of OTU_{97} per supergroup.

Table 2. Taxonomic groups, mainly at class level, ordered follow pyrotags abundance. The respective number of OTUs (97% of similarity) and the average percentage per station is shown.

Group	Pyrotags	OTUs	Average
Collodaria	78394	156	18.288
Chrysophyceae	47864	43	12.243
MALV-II	42421	582	13.100
Basidiomycota	36284	43	10.898
Dinophyceae	26234	377	7.919
Uncertain	23294	200	6.221
Spumellaria	21078	67	6.175
Amoebozoa	14840	27	4.585
Metazoa	14752	98	4.264
Ascomycota	12692	28	3.556
MALV-I	9169	202	2.783
RAD-B	8524	68	2.609
Diplonemea	4303	203	1.244
Acantharea	3632	73	1.132
Bicosoecea	3493	15	1.075
Ciliophora	2293	52	0.796
Larcopyle	2222	13	0.712
MALV-IV	1125	43	0.346
Cercozoa	1040	21	0.342
MAST-1	993	9	0.308
Apusomonadidae	762	3	0.197
Kinetoplastea	679	7	0.222
Choanoflagellida	586	11	0.220
MALV-III	578	23	0.173
Planomonadida	331	4	0.094
Apicomplexa	300	4	0.108
Prymnesiophyceae	248	17	0.079
RAD-C	211	11	0.067
Labyrinthulida	187	12	0.053
MAST-3	105	11	0.033
Embryophyceae	70	3	0.024
Centroheliozoa	67	4	0.019
Picobiliphyta	61	7	0.018
Raphidophyceae	48	1	0.013
Prasinophyceae	40	5	0.011
MALV-V	36	4	0.010
Perkinsea	33	3	0.011
Telonemia	28	7	0.009
MAST-8	23	2	0.006
Bolidophyceae	21	4	0.006
Eustigmatophyceae	20	2	0.006
MAST-4	18	2	0.005
Bacillariophyceae	16	4	0.005
MAST-9	14	4	0.004
Dictyophyceae	13	2	0.004
MAST-7	10	2	0.003
MAST-10	7	1	0.002
MAST-6	2	1	0.001
RAD-A	2	1	0.001

of 82% in station 131) and NADW enriched (14%). They were also contributors in the northern part of CDW (29% in stations 109-121). MALV-II was widespread in all water types, exhibiting a more homogeneous distribution than the other groups (between 11% and 20%). The distribution of Uncertain was not uniform and formed a large share of pyrotag abundance in only three stations

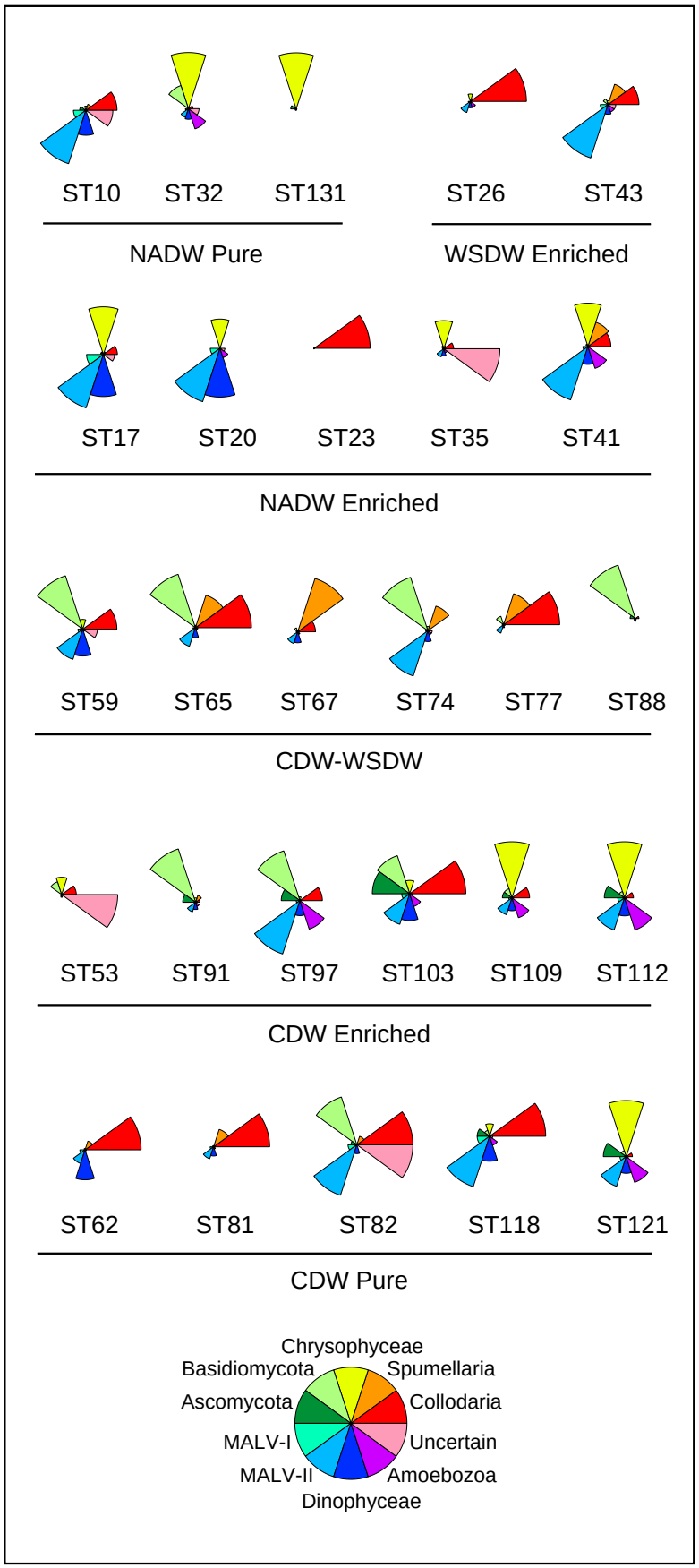


Figure 7. Relative abundance of the ten most abundant phylogenetic groups in all deep samples. Stations are grouped by their respective water mass.

(43%, 41%, 13% in stations 53, 35 and 10, respectively), whereas in the remaining 24 stations were below 4%. The other groups (Amoebozoa, Dinophyceae and MALV-I) showed sporadic peaks in few stations.

Dominant OTUs in the deep ocean

We analyzed the first twenty most abundant OTUs in detail (Table 3). Surprisingly, seven of them had high similarity (>97%) with cultured organisms and were present in almost all stations (at least 25 stations). The most abundant was a colonial Collodaria (45,261 pyrotags) that was 90% similar to *Collophidium ellipsoide*, followed by a Basidiomycota (29,099 pyrotags) and a Chrysophycaea (23,512 pyrotags). The two fungal OTUs in the list were very similar to terrestrial strains (98% to *Tilletiopsis minor* and 99% to *Engyodontium album*). Some species detected (e. g. *Pedospumella encystans* and *Platyamoeba contorta*) are known to have a cyst-stage in their life cycle, indicating the possibility of sampling a dormant (non active) organism. This could explain the fact that a widespread pyrotag (the fifth abundant OTU) was close to the photosynthetic dinoflagellate *Lepidonium chlorophorum* (99% similarity). The first OTU of MALV-II appears in the thirteen position: whereas MALV-II is the most represented group in most samples, not a single OTU appear as dominant, indicating a large diversity and high eveness of this group.

Novel diversity in the deep ocean

A total of 200 OTUs were classified as Uncertain since they were too distant (e-value above 10^{-50} or less than 90% similarity) to any sequence in reference datasets (Table 2). As mentioned before, the distribution of these sequences was not uniform but accumulated in three stations. We visually checked whether or not these pyrotags aligned with reference sequences (therefore having the V4-18S rDNA signature) and the majority did. For each uncertain OTU we identified the closest environmental match (CEM) and the closest cultured match (CCM) in a BLAST search (del Campo and Massana 2011), and OTUs having the same CCM were collapsed in Table S1. Many of these uncertain OTUs could be assigned to large taxonomic categories in base of the CCM hit, forming probably novel lineages within them. In some cases (24%) these were similar to environmental sequences not yet classsified and excluded from reference databases. Most were Rhizaria (86) and Alveolate (41). The most abundant group was Collodaria with 51 OTUs, followed by Dinophyceae with 22 and MALV with 11. Some (12) were metazoan and were not further discussed here. A group of 20 OTUs formed a second level of novelty, since the sequence coverage in the Blast search was below 50% even with the CEM, indicating very low sequence similarity. These were assigned tentatively to the taxonomy of the CCM as Novel-Group name. Finally, a third level of novelty was represented by a group of 27 OTUs with no Blast hit in GenBank. Of these, only 5

Table 3. The twenty most abundant OTUs are showed with their number of tags, occurrence, taxonomic identification and similarity with the closer environmental match (CEM) and the closer cultured match (CCM).

OTU ID	Pyrotags	OCC	Group	CEM	% SI	CCM	% SI
146	45261	27	*Collodaria*	GU825331	90	*Collophidium ellipsoidae*	90
6539	29099	27	*Basidiomycota*	HQ438183	99	*Tilletiopsis minor*	98
941	23512	27	*Chrysophyceae*	JQ782092	99	*Pedospumella encystans*	98
3736	16125	24	*Spumellaria*	EF172914	99	*Cladococcus viminalis*	96
2627	13645	27	*Dinophyceae*	EU500130	100	*Lepidodinium chlorophorum*	99
309	11748	27	*Ascomycota*	GQ120160	99	*Engyodontium album*	99
1730	11131	27	*Amoebozoa*	GU320596	99	*Platyamoeba contorta*	90
2006	8958	20	*Uncertain*	JX194706	77	*Collozoum Serpentinum*	85
2275	8364	25	*Chrysophyceae*	KC306509	98	*Ochromonas distigma*	97
1165	8228	26	*Collodaria*	GU219126	99	*Collophidium ellipsoidae*	94
2418	8151	27	*Metazoa*	AY937332	99	*Gilia reticulada*	98
7646	6784	27	*Chrysophyceae*	HM749946	99	*Mallomonas Tonsurada*	92
2825	5694	27	*MALV-II*	FN598288	100	*Amoebophyra sp.*	89
6489	5597	25	*Uncertain*	GU824572	82	*Collozoum Serpentinum*	88
4675	4604	25	*Chrysophyceae*	KC306509	98	*Ochromonas distigma*	97
3936	4472	21	*Collodaria*	GU825728	96	*Collophidium ellipsoidae*	96
149	4404	20	*Collodaria*	AY046728	96	*Collophidium ellipsoidae*	84
4324	3565	27	*MALV-II*	JX194526	98	*Amoebophyra sp.*	90
5203	3552	15	*Collodaria*	GU824619	82	*Collozoum Serpentium*	94
7437	2568	27	*Amoebozoa*	FN598227	98	*Platyamoeba contorta*	89

appeared more than 10 times and the most abundant only had 229 pyrotags. These were named as True novel, being potentially new high-rank groups not yet described, although it was not even clear if they were true 18S rDNA.

Comparing tag-sequencing and metagenomes

We extracted 18S rDNA sequences from metagenomes (miTags) prepared from the same samples (24 samples in common for metagenomics and pyrosequencing), in order to compare the relative abundance of taxonomic groups inferred by both approaches. Metagenomic data is not constrained by typical PCR biases, and is used as qualitative and semi-quantitative confirmation of the PCR-based pyrotags. Of the 19,434 miTags retrieved, only 3981 showed a similarity above 97% in an alignment of at least 100 bp with the 454-sequencing pool. This precentage (20.5%) is what should be expected by the size of the 454 amplicons (380 bp) with respect by the complete 18S rDNA (21.3%). In general, the percentage of supergroups was very similar by the two approaches (Figure 7a). The most striking differences were a lower representation of Alveolata in miTags than in pyrotags (16.4% versus 25.7%), and the much larger representation of Excavata in miTags (10.7% versus 1.5%).

To compare both approaches at a finer phylogenetic level, we blasted the twenty most abundant OTUs against the pool of miTag sequences using the same similarity criteria (97% in at least 100 bp alignments). Three Chrysophyte OTUs were similar and many miTags affiliated indifferently to one of the three, and this was solved by pooling the relative abundance of these three OTUs in the two approaches. Interestingly, these abundant OTUs recovered a large fraction of miTags.

Thus OTU_146 belonging to Collodaria (45,261 pyrotags) retrieved 829 miTags, OTU_6539 (Basidiomycota, 29,099 pyrotags) retrieved 323 miTags, and the composite of the three chrysophyte OTUs (23,512 pyrotags) retrieved 419 miTags. When expressed as percentage of the respective datasets (normalized per sample), the relative abundance of these 18 OTUs was highly correlated in the two approaches (Figure 7b), with a very high significance ($p<0.0001$), an R^2 of 0.70 and a slope of 0.75.

Discussion

As far as we know this is the first study that analyzes the diversity of microbial eukaryotes assemblages (from 0.8 to 20 µm in size) at the interface between the bathypelagic and the abyssopelagic realm (around 4000 m depth). The geographic coverage in several different basins acquired during the Malaspina-2010 expedition allowed us to make this study at a global scale. Considering the

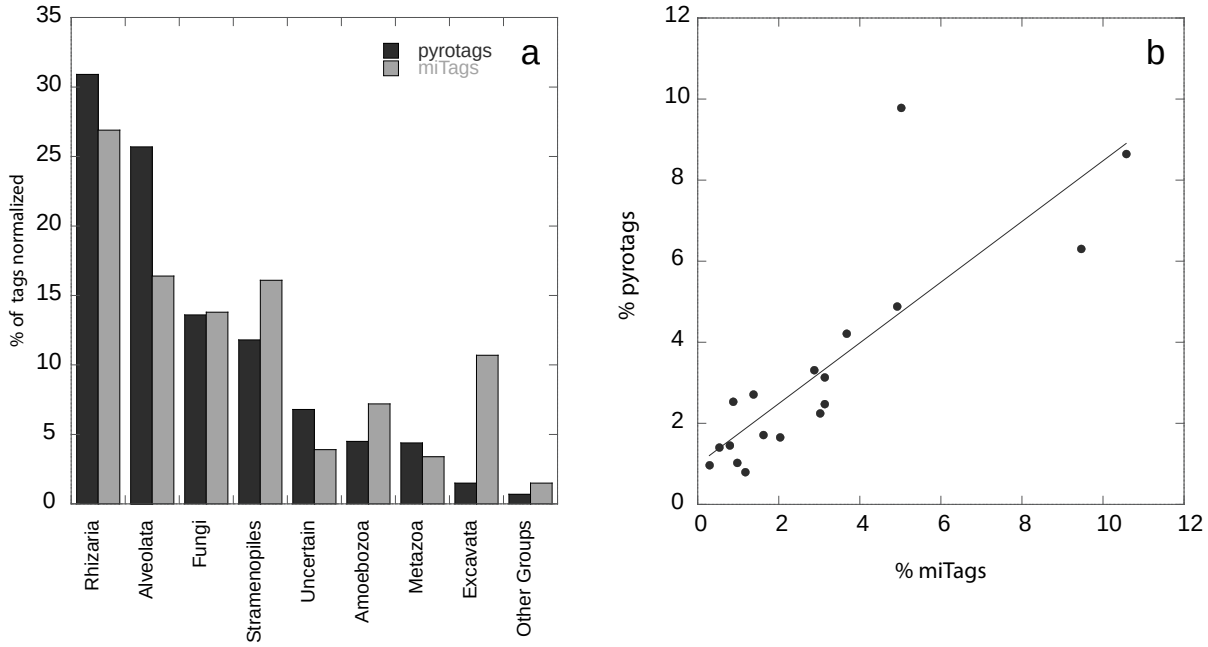


Figure 8. Comparison of deep protist diversity between the pyrosequencing and the metagenomics approaches (**a**) Relative abundance of each supergroup derived from the same 24 stations, averaging in both cases the relative abundance in each station. (**b**) Relationship of the relative abundance of miTags retrieved from the 20 most abundant OTUs.

important role of microeukaryotes in food webs (Massana 2011, Richards *et al.* 2012) our main aim was to characterize the global pattern of their diversity, as a first step for understanding the factors structuring deep microbes and the possible trophic roles they exhibit. Pyrosequencing has proven to be an adequate tool to exhaustively describe the composition of microbial communities (Charvet *et al.* 2012, Logares *et al.* 2012, Kilias *et al.* 2013). Here we report a considerable number of sequences (359,163 pyrotags), well supported in their relative group abundance by a

parallel metagenomic work, which are the base for a comprehensive diversity assessment.

Despite the uniformity of some physical parameters like pressure, temperature and salinity, the deep ocean is not a homogeneous environment, since the concentrations of organic matter and other chemical components display considerable variability (Hansell *et al.* 2009, Hamme and Emerson 2013), which may then shape taxonomic composition. Considered at class level, only a few groups dominate globally deep microeukaryote composition. Thus, Collodaria, Basidiomycota, Chrysophyceae, MALV-II and Dinophyceae account together for 62.4% of the pyrotags from the global deep ocean (Table 2) and always explain the majority of reads in each single individual sample (Figure 1). The relative proportion of these five groups changes dramatically between sites, revealing a high level of heterogeneity, partially explained by abiotic and biotic parameters. Similarly, at the OTU level, we identified 42 OTUs found in all stations and that together account for 50% of pyrotags. We concluded that, despite particular classes and OTUs were widespread in the deep ocean, a typical global deep ocean community could not be identified, due to the large variation in their relative abundances.

Considered each sample alone, the rarefaction curve did not show a sign of saturation, suggesting that the 454-pyrosequencing applied here was not sampling exhaustively the microeukaryotes diversity (Figure 3a). Shannon indices, with a global average of 3.6 (Figure 3b), did not show a regular pattern along the cruise nor a relation with environmental parameters. This averaged value was largest than the one found for microeukaryotes in an ice-covered lake (0.69- 2.18, Bielewicz *et al.* 2011) and similar to values found in surface marine samples (2.66-9.55, Kok *et. al.* 2012). Despite these values seem to suggest that deep ocean microeukaryotes are as diverse as surface ones, it is important to highlight that the use of different protocols, most notably the sampling size, can affect Shannon indices.

The geographic distance does not either explain very well the degree of genetic similarity among samples (Figure 5), and only a few couple of close stations were very similar (e.g. 17 and 19; 112 and 121). This is in agreement with other studies that found similar communities of microeukaryotes at spatial scales of thousands of kilometers (Scheckenbach *et al.* 2010, Salani *et al.* 2012). This similarity is caused probably by similar environmental conditions (summarized by the water masses properties) that drive community structure. Indeed, similar water masses commonly occupy large distances (thousands of km). Interestingly, considering the stations between 3000 and 4000 m, the most important factor explaining microeukaryotes composition was the water mass type, which explained 28% of the variability among stations (Figure 4a, Table 1) and, in the presence of a biotic parameter (ratio Prok:HP), still explained 26%. This could be due to conspicuous

differences in chemical composition of the water mass, or to its particular history. The link between community structure of marine prokaryotes and water masses has been reported previously (Agogué *et al.* 2008, Varela *et al.* 2008, Galand *et al.* 2009, Kirchman *et al.* 2010). Our data seems to agree with the general view of biogeography of microbes obtained from surface waters (Massana and Logares 2013): there is not a clear geographical restriction for microbial dispersal, and the environment clearly selects for specifically-adapted microbes.

Surprisingly, in a world dominated by heterotrophy, the abundance of bacteria, which could be seen as the main food source for deep microeukaryotes, did not have a direct effect on taxonomic composition (Table 1). However, the situation changed dramatically when considering the ratio prokaryotes to microeukaryotes. In a subset of 20 stations this ratio explained 34% of the variability in the community structure with high significance. In this case, the biotic component has more weight than the abiotic parameters in structuring the community of microeukaryotes. The average bathypelagic ratio (ca. 5000) was rather high as compared with surface samples and, particularly for the samples with the highest ratio, questions the importance of bacterivory in the deep ocean ecosystem. Despite different sign of the importance of bacterivory (Pernice et al. submitted) it is likely that, in the dark ocean, heterorotrophy is present also in the form of osmotrophy and parasitism. Generally, these alternative heterotrophic processes are linked to particular taxonomic classes, so a fine taxonomic analysis is required to understand the functioning of the deep ocean system.

There is only a handful number of molecular surveys of deep marine protists, done by amplifying 18S rDNA genes with universal primers and Sanger sequencing in samples up to 3000 m. In these studies, deep protists appeared dominated by Alveolates, mainly MALV-II (Lopez-Garcia *et al.* 2001, Not *et al.* 2007) or radiolarians (Countway *et al.* 2007, Sauvadet *et al.* 2010). These groups were also very important in our survey, but here we also added Chrysophyte and Fungi. Fungi have been reported to dominate in deep sediments in base of clone libraries (Edgcomb *et al.* 2011a) or of 454-pyrosequencing surveys based on rRNA (Orsi *et al.* 2013), but as far as we known our study is the first identifying Fungi as a dominant group in deep water samples. Moreover, the presence of chrysophytes has never been pointed before in deep samples.

In our global survey, the most abundant class was Collodaria (18.3% of the reads on average). Collodaria is an order of Polycystinea, a class that includes mostly species with colonial lifestyle and without silification (Ishitani *et al.* 2012). In surface, Collodaria persistently bear photosynthetic endosymbionts and are ecologically categorized as protists with phototrophic behavior (Stoecker *et al.* 2009). Therefore, it is intriguing which type of protists are these deep collodaria. Surprisingly Chrysophytes, generally represented by few sequences in surface marine waters (Massana

and Pedrós-Alió 2008, del Campo and Massana 2011), is the second most abundant class. These protists can be phagotrophs or facultative or strict osmotrophs (Holen and Boraas, 1996). A few cultured isolates (*Pedospumella encystans, Ochromonas distigma, Paraphysomonas bandaiensis*) explain 79% of chrysophyte sequences (considering a similarity >97%). The MALV-II is a ribogroup belonging to Alveolates, defined for the first time by Lopez-Garcia *et al.* in 2001 in samples from deep water, and also widely found in surface waters. It is known that this ribogroup contains the genus *Amoebophyra*, a parasite of dinoflagellates (Coats and Park 2002), so parasitism could be extended to the entire group (Massana 2011), and perhaps having other hosts. Their high relative-abundance can be partially due to multiple copies of the rDNA operon, but still the prevalence of MALV-II suggests an unanticipated role of parasitism in deep waters. Interestingly in Atlantic waters there was a good positive correlation (R^2=0.96, p=0.0005) between the relative tag abundance of MALV-II and Metazoans, which could be a signal of metazoans being infected by MALV-II parasites. A striking finding was the abundance of fungi, particularly Basidiomycota. Commonly discharged from clone library surveys, fungi, which are widespread in different environments, are gaining attention in the last few years. They are important in deep sea sediments (Edgcomb *et al.* 2011a) and plankton, where fungal diversity is dominated by Basidiomycota and Ascomycota with a probable yeast life-style (Bass *et al.* 2007, Richards *et al.* 2012). Finally, considering a potential PCR bias of our V4 primers against Excavata (Pawlowski *et al.* 2011), we probably undersampled this Supergroup. Excavata, which include many bacterivorous forms, are typically more abundant in deep waters than at surface (Lara *et al.* 2009), representing until 15% of FISH-counted cells in deep protistan assemblages (Morgan-Smith *et al.* 2013). In fact, they reach 11% of total reads in the parallel metagenomics approach. With this percentage Excavata cells could be very important phagotrophs in the deep ocean.

Compared with the better known surface microeukaryotes (Massana 2011), the first obvious difference in deep samples is the virtual absence of photosynthetic groups. Some sequences from these groups (Bacillariophyta, Bolidophyta, Dictyochophyta, Prasinophyceae, Prymnesiophyceae, Raphidophyta, Eustigmatophyceae) have been indeed retrieved in the deep ocean, but together they only represent 0.14% of the pyrotags and 2% of the OTUs. The presence of phototrophic signal in bathypelagic waters is likely due to sinking particles, athough we do not totally dismiss the idea of some species being facultative heterotrophs. The absence of photosynthetic groups affects the importance of Stramenopiles in the global deep community, while Alveolate and Rhizaria are represented with percentages similar as in surface. Surprisingly, in a world dominated by heterotrophy, the MASTs ribogroups that are important phagotrophs in the surface ocean (Massana *et al.* 2014), are poorly detected. For instance, one of the most abundant and widely distributed

ribogroups, MAST-4 (Rodríguez-Martínez *et al.* 2009), is unrepresented in our data-set, with only 2 OTUs and 18 pyrotags. The difference in the composition of taxonomic groups between surface and deep waters implies different players in the flux of energy and carbon through the microbial loop.

The amount of pyrotags of Uncertain identification was 6% (Figure 5), and these were concentrated in three hotspots (stations 35, 53 and 82), one in each ocean (Figure 6). Together, these three samples represented 66% of the Uncertain sequences. This could be explained by the geographical peculiarity of these stations, which are near the bottom (35), close to hydrothermal vents (82) or belonging to an isolated basin (53). The relative abundance of uncertain-groups parallels the abundance of taxonomical groups, being most uncertain tags related to Rhizaria and Alveolata (Table S1). Considering the peculiarity of the deep ocean environment and the sampling effort, the percentage of novel species discovered is lower than expected, suggesting a certain degree of genetic similarity between deep ocean species and the surface ones.

Considering the prokaryotes to microeukaryotes cell abundance ratio (Figure 4b) it is possible to propose a tentative scenario for the deep ecosystem functioning, assigning a possible trophic role to the four principal groups (Collodaria, Chrysophytes, Fungi, MALV-II). When this ratio is low, similar to the typical values in surface communities regulated by bacterivory (between 1000 and 2000; Pernice *et al.* submitted), the diversity is strongly dominated by Collodaria, suggesting a possible phagotrophic role for this class. At ratios around the average bathypelagic values (from 2501 to 4950) Collodaria tend to decrease in favour of Chrysophytes, which are also good candidates to be deep grazers. In fact, the most abundant Chrysophyte OTU (representing 45% of chrysophyte pyrotags, Table 3) is 98% similar to *Pedospumella encistans,* a proved bacterivore. Values of ratio higher than the bathypelagic average could be caused by a decrease of the impact of bacterivory, suggesting communities where osmotrophy starts to have a role. Thus, Fungi, which have a proved osmotrophic lifestyle (Richards *et al.* 2012), start to appear with ratios higher than ca. 5000. The presence of Fungi could partially explain the constant decrease in DOC from the Southern Ocean until the North Pacific (Hansell *et al.* 2009). Interestingly, Chrysophytes and Fungi were never abundant in the same sample (except in station 32 in the Atlantic Ocean). The high ratio communities (light blue points, Figure 4b) were not exclusively dominated by Fungi and it is worth to remember that the ratio explains only 34% of the variability between samples. Despite been globally the most abundant group, MALV-II was never found to dominate a community. Taking into account that they are probable parasites, they do not compete for the same resources with the other major groups. Parasitism could be also a trophic alternative for Fungi, as suggested by the most abundant fungal OTUs (Table 3), but this is a mechanism not fully understood nor

deeply investigated here.

The present study allowed to describe the general pattern in the diversity of microbial eukaryotes in the deep ocean. Thanks to the magnitude of Malaspina-2010 expedition we collected a large number of tags from separate marine regions that were further validated by a parallel metagenomic survey. From this well curated dataset we identified few groups (Basidiomycota, Collodaria, Chrysophytes and MALV-II) that alternatively dominate communities, whose similarities were partially explained by the water mass they belong and by the ratio in the abundance of prokaryotes to microeukaryotes. The fact that Chrysophytes and Fungi dominated some deep water samples was never observed before. Our research sets the global architecture of the deep ocean protistan communities and is the essential first step for a more detailed investigation of such an interesting environment.

Acknowledgements

This study was supported by the Consolider-Ingenio Malaspina 2010 financed by the former Ministry of Science and Innovation (MICINN). We thank the scientists that sampled for DNA in the different legs of the cruise: Guillem Salazar, Francisco Cornejo, Cristina Diéz, Elena Lara, Encarna Borrull and Dolors Vaqué.

References

Acinas S, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz M (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71:** 8966.

Agogué H, Brink M, Dinasquet J, Herndl GJ (2008). Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. *Nature* **456:** 788-791.

Angel MV (1993). Biodiversity of the Pelagic Ocean. *Conserv Biol* **7:** 760-772.

Azam F, Fenchel T, Field J, Gray J, Meyer-Reil L, Thingstad F (1983). The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* **10:** 257-263.

Bass D, Richards T, Matthai L, Marsh V, Cavalier-Smith T (2007). DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* **7:** 162.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Bielewicz S, Bell E, Kong W, Friedberg I, Priscu JC, Morgan-Kiss RM (2011). Protist diversity in a permanently ice-covered Antarctic Lake during the polar night transition. *ISME J* **5:** 1559-1564.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Meth* **7:** 335-336.

Charvet S, Vincent WF, Comeau AM, Lovejoy C (2012). Pyrosequencing analysis of the protist communities in a High Arctic meromictic lake: DNA preservation and change. *Front Microbiol* **3**: 422. doi: 10.3389/fmicb.2012.00422

Clarke KR, Ainsworth M (1993). A method of linking multivariate community structure to environmental variables. *Mar Ecol Prog Ser* **92**: 205-219

Coats DW, Park MG (2002). Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophyra* (Dinophyta): parasite survival, infectivity, generation time and host specificity. *J Phycol* 38: 520-528.

Countway PD, Gast RJ, Dennet MR, Savai P, Rose JM, Caron DA (2007). Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ Microbiol* **9:** 1219-1232.

Del Campo J, Massana R (2011). Emerging Diversity within Chrysophytes, Choanoflagellates and Bicosoecids Based on Molecular Surveys. *Protist***:** 1-14.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200.

Edgcomb V, Kysela D, Teske A, de Vera Gomez A, Sogin M (2002). Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc Natl Acad Sci USA* **99:** 7658 - 7662.

Edgcomb V, Beaudoin D, Gast R, Biddle JF, Teske A (2011a). Marine subsurface eukaryotes: the fungal majority. *Environ Microbiol* **13:** 172-183.

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C *et al* (2011b). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J***:** 1-13.

Finn RD, Clements J, Eddy SR (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39:** 29-37.

Fuhrman J, McCallum K, Davis A (1992). Novel major archaebacterial group from marine plankton. *Nature* **356:** 148-149.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *PNAS* **106** (52): 22427-22432.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L *et al* (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41:** 597-604.

Hansell, D.A., C.A. Carlson, D.J. Repeta, and R. Schlitzer. 2009. Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography* **22**(4):202–211

Hamme RC, Emerson SR (2013). Deep-sea nutrient loss inferred from the marine dissolved N2/Ar ratio. *Geophys Res Lett* **40:** 1149-1153.

Holen DA, Boraas ME (1996) Mixotrophy in chrysophytes. In: D Craig, CD Sandgren, JP Smol, J Kristiansen (Eds.) *Chrysophyte algae. Ecology, phylogeny and development*, University Press, Leiden (1996), pp 119–140.

Ishitani Y, Ujiié Y, de Vargas C, Not F, Takahashi K (2012). Phylogenetic Relationships and Evolutionary Patterns of the Order Collodaria (Radiolaria). *PLoS ONE* **7:** e35775.

Kilias E, Wolf C, Nöthig E-M, Peeken I, Metfies K (2013). Protist distribution in the Western Fram Strait in summer 2010 based on 454-pyrosequencing of 18S rDNA. *J Phycol* **49:** 996-1010.

Kirchman DL, Cottrell MT, Lovejoy C (2010). The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* **12:** 1132-1143.

Kok SP, Kikuchi T, Toda T, Kurosawa N (2012). Diversity and community dynamics of protistan microplankton in Sagami Bay revealed by 18S rRNA gene clone analysis. *Plankton Benthos Res* **7**(2):75-86

Lara E, Moreira D, Vereshchaka A, López-García P (2009). Pan-oceanic distribution of new high-

ly diverse clades of deep-sea diplonemids. *Environ Microbiol* **11:** 47-55.

Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana R (2012). Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J* **6:** 1823-1833.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H *et al* (2013). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol*.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603 - 607.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L *et al* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Massana R, Pedrós-Alió C (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11:** 213-218.

Massana R (2011). Eukaryotic picoplankton in surface oceans. *Annu Rev Microbiol* **65:** 1-47.

Massana R, del Campo J, Sieracki ME, Audic S, Logares R (2014). Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J*. **8:** 854-866

Massana R, Logares R (2013). Eukaryotic versus prokaryotic marine picoplankton ecology. *Environ Microbiol* **15:** 1254-1261.

Michaels AF, Silver MW (1988). Primary production, sinking fluxes and the microbial food web. *Deep Sea Res Pt I* **35:** 473-490.

Morgan-Smith D, Herndl GJ, van Aken HM, Bochdansky AB (2011). Abundance of eukaryotic microbes in the deep subtropical North Atlantic. *Aquat Microb Ecol* **65:** 103-115.

Morgan-Smith D, Clouse MA, Herndl GJ, Bochdansky AB (2013). Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. *Deep Sea Res Pt I* **78:** 58-69.

Nagata T, Tamburini C, Arístegui J, Baltar F, Bochdansky AB, Fonda-Umani S *et al* (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. *Deep Sea Res Pt II* **57:** 1519-1536.

Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ (2007). Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. *Environ Microbiol* **9:** 1233-1252.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB *et al* (2013). *Vegan: Community Ecology Package*. *R package version 2.0-7*. http://CRAN.R-project.org/package=vegan

Orsi W, Biddle JF, Edgcomb V (2013). Deep Sequencing of Subseafloor Eukaryotic rRNA Re-

veals Active Fungi across Marine Subsurface Provinces. *PLoS ONE* **8:** e56335.

Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, Amaral-Zettler L *et al* (2011). Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. *PLoS ONE* **6:** e18169.

Pernice MC, Logares R, Guillou L, Massana R (2013). General patterns of diversity in major marine microeukaryote lineages. *PLoS ONE* **8:** e57170.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** 590-596.

Richards TA, Jones MDM, Leonard G, Bass D (2012). Marine Fungi: Their Ecology and Molecular Diversity. *Annu Rev Mar Sci* **4:** 495-522.

Rodríguez-Martínez R, Labrenz M, Del Campo J, Forn I, Jürgens K, Massana R (2009). Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. *Environ Microbiol* **11:** 397-408.

Salani FS, Arndt H, Hausmann K, Nitsche F, Scheckenbach F (2012). Analysis of the community structure of abyssal kinetoplastids revealed similar communities at larger spatial scales. *ISME J* **6:** 713-723.

Sauvadet A-L, Gobet A, Guillou L (2010). Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environ Microbiol* **12:** 2946-2964.

Scheckenbach F, Hausmann K, Wylezich C, Weitere M, Arndt H (2010). Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. *PNAS* **107:** 115-120.

Stoeck T, Taylor G, Epstein S (2003). Novel eukaryotes from a permanently anoxic Cariaco Basin (Caribbean Sea). *Appl Environ Microbiol* **69:** 5656 - 5663.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W *et al* (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19:** 21-31.

Stoecker D, Johnson M, deVargas C, Not F (2009). Acquired phototrophy in aquatic protists. *Aquat Microb Ecol* **57:** 279-310.

Suzuki R, Shimodaira H (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22:** 1540-1542.

Vaqué D, Gasol JM, Marrasé C (1994). Grazing rates on bacteria: The significance of methodology and ecological factors. *Mar Ecol Prog Ser* **109:** 263-274.

Varela MM, Van Aken HM, Herndl GJ (2008). Abundance and activity of Chloroflexi-type SAR202 bacterioplankton in the meso- and bathypelagic waters of the (sub)tropical Atlantic. *Environ Microbiol* **10:** 1903-1911

Wessel P, Smith WHF, Scharroo R, Luis J, Wobbe F (2013). Generic Mapping Tools: Improved Version Released. *Eos, Trans Am Geophys Union* **94**: 409-410.

Supplementary material

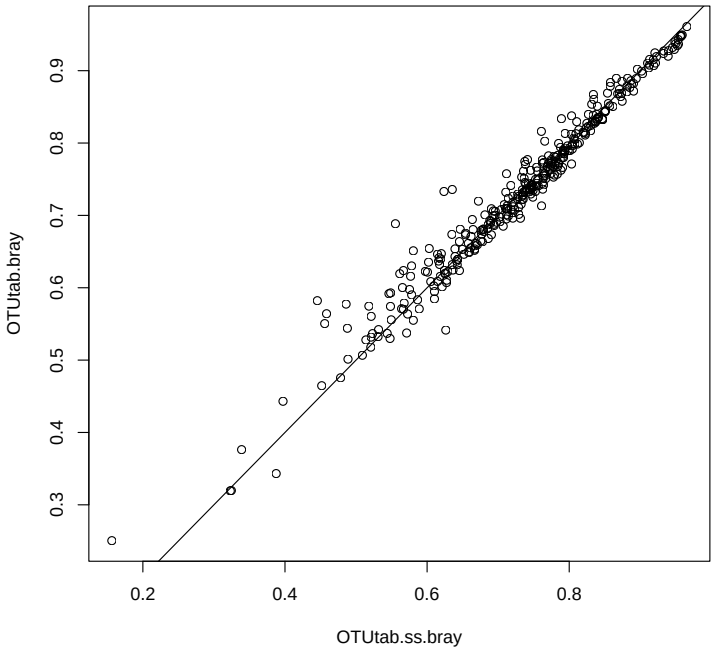


Figure S1 Mantel test relating Bray-Curtis genetic distances obtained by the total pool of pyrotags in the initial OTU table and the subsampled OTU table (at 6625 pyrotags per sample). The subsampled dataset fits very well with the original one.

Table S1. Uncertain sequences, tags with similar closer cultured match (CCM) were grouped. The table shows the probably taxonomy and for each group the range of similarity (%) with cultured and environmental sequences (CEM), the asterisk marks a coverage <50%.

CCM	Taxonomy	OTUs	Pyrotags	CEM %	CCM %
Acantharia sp. (JX661016)	Acantharea	1	2	83	82
Acanthocyrta haeckeli (JN811157)	Acantharea	1	14	97	79
Acanthostaurus purpurascens	Acantharea	1	2	100	82
Heteracon biformis (KC172870)	Acantharea	2	44	99	87-88
Stauracon pallidus (KC172869)	Acantharea	1	2	90	82
Staurolithium sp. (JN811182)	Acantharea	1	8	83	83
Linguamoeba leei (AY183886)	Amoebozoa	2	32	82-86	86
Vanella sp. (AY929908)	Amoebozoa	3	244	83-87	85-87
Psychodiella sp. (GQ329865)	Apicomplexa	1	11	88	96 *
Eunotogramma sp. (JN975245)	Bacillariophyta	1	2	96	89
Adriamonas peritocrescens (GQ911660)	Bicosoecid	1	19	94	85
Gromia oviformis (AJ457811)	Cercozoa	3	11	86-89	80*-95*
Helkesimatix sp.	Cercozoa	1	2249	-	90
Salpingoeca sp.(DQ995808)	Choanoflagellate	1	25	86	81
Caryotricha minuta (EU275202)	Ciliophora	1	10	96	94
Eulotes eurhyalinus(EF094967)	Ciliophora	1	23	91	95
Homologastra setosa (GU590870)	Ciliophora	1	122	90	89
Protocruzia sp. (JF694044)	Ciliophora	1	3	84	88
Tetrahymena corlissi (U17356)	Ciliophora	1	2	81	78
Trachelocerca sagitta (KC542935)	Ciliophora	1	2	96	77
Uroychia binucleata (EF198667)	Ciliophora	1	7	80	80
Ceratospyris hiperborea(HQ651791)	Collodaria	2	257	83-86	82-83
Collophidium ellipsoides (AB690557)	Collodaria	15	724	84-99	83-89
Collozoum Serpentinium(AF018162)	Collodaria	34	17321	77-86	77-95*
Developayella sp. (JX272636)	Collodaria	1	2	84	92*
Amphidinium semilunatum (AF274256)	Dinoflagellate	1	16	98	83
Azadinium sp. (JQ247707)	Dinoflagellate	2	13	95-98	81-86
Gymnodinium sp. (JQ639760)	Dinoflagellate	3	18	87-96	78-81
Gyrodinium sp. (AB120002)	Dinoflagellate	2	5	89-95	78-89
Lepidonium sp.(AB686255)	Dinoflagellate	1	22	99	80
Peridinium cinctum (EF058245)	Dinoflagellate	1	8	94	81
Pfiesteria piscicida (FJ600090)	Dinoflagellate	1	10	94	81
Phalacroma mitra (AB551248)	Dinoflagellate	1	6	95	78
Polykrikps kofoidii (DQ371292)	Dinoflagellate	1	3	89	81
Porocentrum sp. (AY551272)	Dinoflagellate	4	31	93-98	81-90
Scrippsiella trochoidea (HM483396)	Dinoflagellate	2	9	86-89	77-84
Spniferodinium sp (AB626150)	Dinoflagellate	1	7	96	82
Symbiodinium pilosum (X62650)	Dinoflagellate	1	20	92	83
Woloszynskia cincta (FR690459)	Dinoflagellate	1	3	87	78
Diplonema sp. (AY425011)	Diplonemids	1	4	97	76
Petalomonas sphangonophila (GU477297)	Euglenida	1	4	77	86
Bodo saltans (DQ207571)	Kinetoplastida	1	2	81	81
Thrausthochytridae sp (AY872261)	Labryrinthulidae	2	57	97-99	92-95
Amoebophyra sp. (AF472555)	MALV	10	130	82-99	84-85
Euduboscquella sp.	MALV	1	21	99	78
e.g. Drawida sp. (HQ728930)	Metazoa	12	84	76-98	74-97
Mataza hastifera (AB558956)	Cercozoa	11	113	82-88	92*-95*
Korotnevella stella (AY686573)	NOVEL-Amoebozoa	2	144	82 *-83*	80*-81*
Lesquereusia spiralis (JQ519506)	NOVEL-Amoebozoa	2	41	87 *-88*	87 *
Platyamoeba contorta (DQ229953)	NOVEL-Amoebozoa	1	34	94*	95*
Squamamoeba japonica (JN638031)	NOVEL-Amoebozoa	2	56	84*-89*	85
Vanella sp. (AY929904)	NOVEL-Amoebozoa	1	34	92*	91*
Psychodiella sp. (GQ329865)	NOVEL-Apicomplexa	1	18	87*	96*
Salpingoeca sp.(DQ995808)	NOVEL-Choanoflagellate	2	11	86*-88*	80-81
Acrobeloides maximus (EU306344)	NOVEL-Metazoa	1	15	78*	79*
Bentheogennema intermedia	NOVEL-Metazoa	1	3	98*	98*
Candacia columbiae(AB625974)	NOVEL-Metazoa	2	25	96*	96*-99
Gyrodactylus colemanensis(JF836090)	NOVEL-Metazoa	1	3	95*	90*
Monstrilla clavata	NOVEL-Metazoa	1	9	95*	97*
Chrysocrumulina sp.	NOVEL-Prymnesiophyceae	1	2	92*	92*
Blasocystis sp (KC148211)	NOVEL-Stramenopiles	1	66	87 *	91 *
Odontonella aurita(HQ912688)	NOVEL-Stramenopiles	1	2	98*	95*
Sticholonche sp. (HQ651785)	RAD_B	2	10	94-98	82-83
Haplosporidian sp (AY449716)	Rhizarian	4	18	83-92	74-96*
Paradinium poucheti (EU189031)	Rhizarian	2	7	87-99	82-90*
Spongocore puella (AB617587)	Spumellaria	1	74	99	82
Sphaerozoum sp. (AB690556)	Spumellaria	4	321	86-90	80-90
True novel	True novel	27	677	-	-

Synthesis of results and general discussion

The first objective of this thesis was to improve the information about the genetic structure of microeukaryotes groups, mainly at class-rank level, with the final goal of building a database of well-curated and reliable sequences. Once defined this reference sequence dataset, employing only molecular surveys done by Sanger sequencing, it was used as a framework for studying the global diversity of microeukaryotes in the deep ocean using high-throughput sequencing. The diversity of deep microeukaryotes plus the data collected on their abundance and biomass, allowed drawing a refined picture of the bathypelagic environment.

The molecular taxonomy challenge: tools to define a group

The V4 region of 18S rDNA: the best compromise

The management of a large number of sequences that can be retrieved in molecular surveys needs a great initial effort in order to establish good working criteria. Sequences need to be organized in categories (OTUs) defined by given levels of similarity in order to get quantitative values of diversity. The first step of our work was defining the target region within the 18S rDNA, since no sequencing technology is currently ready to analyze the complete gene. In particular the choice was between V4 and V9 regions. This debate was born inside "BioMarKs", an European research project that studies protistan diversity with molecular and microscopical techniques. Seminal studies using 454 pyrosequencing focused on the V9 region (Amaral-Zettler *et al.* 2009; Cheung *et al*. 2009, Stoeck *et al*. 2009), which is a very short region (up to 150 bp) and was optimal for the sequencing technology at that time. With technical advances yielding longer reads (currently up to 400 bp but longer sequences are expected in the near future) it was logical to analyze longer hypervariable regions, such as the V4. We contributed to this discussion by looking for the region that better represented the entire gene. Our results suggested that the variability detected in the V4 region (around 500 bp) was a good indicator of the variability that would be seen analyzing the entire gene. The respective plots (Figure 1, Chapter 2) showed that the slopes of the regression lines between the distances calculated with the entire gene and the V4 region were around 1.4 for three different Supergroups, so distances calculated with the V4 region could be translated to distances with the entire gene by dividing by this value.

How to define a taxonomic class: from phylogeny to clustering

Sequencing technology improves fast, maybe faster than our ability to manage it. Seminal molecular cloning methods typically yielded between 100 and 500 sequences per sample. Despite this number is very low compared with the output of advanced high-throughput sequencing methods,

the overall process is more controlled and the sequences obtained are longer. Thus, the output of clone libraries is very useful to build a reliable alignment that is the base for a correct phylogeny, as we have done in the first chapter (Figure 1). During the "age of clone libraries" building a tree was the best way to define a taxonomic group but nowadays the superproduction of sequences makes this operation more difficult, particularly aligning a great number of genetically distant sequences. Certainly, short sequences can not resolve the taxonomic relationships among distant lineages. In the "age of pyrosequencing" there is a progressive skip from phylogenetic trees to clustering approaches by which sequences are grouped in taxonomic units based on sequence similarity. Grouping of sequences in a tree is based on patristic distances between sequences (branch lengths), which depend on the number and variability of sequences considered and compares each sequence and all the rest. In contrast, the OTU clustering is based on similarity (or corrected genetic distances such as Jukes Cantor) between all pairs of sequences. The absolute value of similarity or distance is independent from the number of sequences considered, since only pairwise comparisons are performed. The scientific community is adopting as routine the clustering of sequences based on similarity, the problem is that in this process the shape of the tree is lost, so now is more difficult to identify outlier and fast-evolving lineages. In the first chapter we tested if the number of OTUs from the two approaches was different, and we found that at distances up to 0.10, the grouping using JC distances (almost equivalent to similarity) or patristic distances (from the phylogenetic tree) showed good correspondence (Figure 2, Chapter 1). Considering that often the clustering is done at 95 to 97% similarity, this is an acceptable result. Despite patristic distances would result in a more accurate and evolutionary robust clustering, similarity clustering is used with 454 datasets, since the alignment step is skipped and it provides a simple and intuitive way to analyze a high number of sequences.

Setting the limits of taxonomic groups

Working with clustering methods it is essential to answer two questions: what is the cut-off that defines an OTU with useful biological meaning (i.e. a species) and what is the maximal distance that can be found within a given group. About the first question, different authors have proposed different cut-off levels (Worden 2006, Jeon *et al*. 2006, Caron *et al*. 2009) but there is little support to justify each hypothetic level, and it would be important that the scientific community arrives at some conclusion into this direction. Due to intragenomic polymorphisms (Introduction, Table 1) and low-frequency sequencing errors we think it is not advisable to use 100% similarity as the OTU definition. A value between 97-99% similarity appears to be a more reasonable criterion because it is high enough to be rather strict, but not so high as to separate sequences only due to intragenomic polymorphism or sequencing errors. In the chapter two we addressed the second

question for groups roughly corresponding to taxonomic classes in classical systematics, pushed by the lack of the graphical output of the tree that made necessary an alternative way to rapidly identify outlier sequences. We performed trees to be sure about the affiliation of the sequences, which were later analyzed clustering group by group. We found that 75% of the class-rank groups had a corrected maximum pairwise genetic distance below 0.25. This is now our general reference to the maximum distance allowed within a class, and it is useful value to interpret the taxonomic equivalence of environmental ribogroups.

The importance of a curated reference database

A good reference database is an essential tool for any molecular study based on short sequences, but current databases target only prokaryotes (GreenGenes) or give a less accurate treatment to eukaryotes (SILVA). The main problem of SILVA is that often lacks of good taxonomic assignations for eukaryotic sequences, especially for the newly discovered ribogroups. A new tool for the identification of eukaryotic 18S rDNA has been published very recently, the PR2 database (Protist Ribosomal Reference Database, Guillou *et al*. 2013). This tool did not exist at the beginning of this thesis. In this frame, I want to highlight the importance of the pool of well-curated sequences from chapter two (8291) that constitutes the core, improved with PR2, of an in-house reference database (MAS9013), which has been used for taxonomic assignation and chimera detection in the second part of the thesis and in other publications in preparation.

Typical composition of epipelagic microeukaryotes

The high-rank diversity observed in chapter 1 in terms of relative abundance of specific lineages (Figure 1), is the typical found in other molecular surveys of marine picoeukaryotes (Massana and Pedrós-Alió 2008, Vaulot *et al*. 2008) and resembles the abundance of groups in chapter 2 (Table 1). Alveolates, mainly MALV-I and MALV-II, dominated the community and represented 47% of the clones, followed by Stramenopiles (19%) and Rhizaria (13%). Fungi were not considered in the first two chapters, since generally they are little represented in the epipelagic environment, globally they are less than 1% of the sequences in clone libraries (Massana and Pedrós-Alió 2008). Differences in the taxonomic composition of epipelagic and deep microeukaryotes, even at high rank clustering levels, are evident (Figure 6, Chapter 4) and will be analyzed in the second part of this discussion. Interestingly, at low distances (minimal 1300 Km), samples strongly differed when analyzed by clone libraries (Figure 6, Chapter 1), and at that time this was explained by undersampling. However, at comparable distances and with a major sequencing effort, we see that differences among deep samples are still present (Figure 5, Chapter 4). This is a clear signal of the strong effect of the environment on community selection.

The deep ocean

Counting microeukaryotes: the need of flow-cytometry

As far as we know this is the first study that applies flow-cytometry, together with microscopy, on a large scale investigation of the abundance of microeukaryotes (Chapter 3). Epifluorescence microscopy is time consuming and is prone to errors of the operator, while flow-cytometry presents other types of problems. Whereas it is possible to identify several populations of photosynthetic microeukaryotes thanks to their pigments (Olson *et al.* 1985; Li *et al.* 1994; Marie *et al.* 2001), to detect heterotrophic microeukaryotes a general stain (in this case SYBR Green) is required. However, this does not discriminate between prokaryotes and eukaryotes, and a continuum between large bacteria and small eukaryotes exists. To solve this problem, we used epifluorescence microscopy on selected samples to set the counting gate in the cytometry software (then applied to the complete dataset) and to check stations with unrealistic abundances. An alternative method, to save time and recover the size and shape of the cells, would be automatic epifluorescence microscopy, which will be implemented in the future in our lab. The comparison of flow-cytometry and microscopic counts (Figure 2, Chapter 3) was very good (R^2=0.82, p<0.0001). Therefore, a large number of samples were then processed by flow-cytometry. The abundance of microeukaryotes thus determined was one of the two parameters in the description of the deep ocean global community.

General features of microeukaryotes in the bathypelagic ocean

Considering the bathypelagic region (1000 to 4000 m), also the focus of the diversity study, the abundance of microeukaryotes averages 14 cells mL^{-1}. This concentration is not constant, and this is particularly evident in the South Pacific, where there is a peak of 58 cells mL^{-1} in the deepest sample of station 98. Regarding cell size structure, the percentage of very small cells (equivalent diameter <3 μm) decreases with depth (Figure 6, Chapter 3) and from the pictures taken for biomass measurements, we know that some of these cells are clearly flagellated. The average biomass of microeukaryotes is 50 pg C mL^{-1} in the 1400 to 4000 m layer. For the diversity study of the deepest samples (~4000 m), we filtered 120 L, meaning that we collected about 1,320,000 cells per sample. Most of the sequences retrieved belong to Rhizaria, followed by Alveolata and Stramenopiles (Figure 6, Chapter 4). At a local level and at lower taxonomic rank, communities are dominated basically by three classes (Collodaria, Chrysophytes, Basidyomicota) and one ribogroup (MALV-II). Formally Collodaria is an *order*, but considering the value of its maximal distance retrieved in the first chapter (Table 1) and that Polycystinea is not a monophylethic group

(Figure 3b, Chapter 2) it was regarded as a *class*. Differences in microeukaryotic abundance and diversity among deep samples are clearly related both to abiotic (oxygen, temperature) and biotic (prokaryotic and viral abundance) parameters.

Culturing bias and deep protists

Surprisingly the analysis of the first twenty OTUs that represent 50% of total reads show several examples of high similarity with cultured organisms (Table 3, Chapter 4). This fact was previously reported for Fungi (Bass *et al*. 2007, Richards *et al*. 2012) but was unexpected for the other groups. For example three OTUs that explain 79% of the chrysophyte tags are very similar to cultured species. We know that in the epipelagic environment there is normally little agreement between the diversity retrieved in molecular surveys and in culture-based surveys (Massana *et al.* 2004a), and considering the environmental characteristics of the dark ocean and its relative isolation we expected to find even less agreement. However, one third of the most abundant OTUs are more than 97% similar with a cultured species and represent totally 28% of the tags. Moreover two of these OTUs are 99% similar to cultivated species. Considering that most cultured species derive from epipelagic samples, we can deduce that at least one quarter of the tags are shared between surface and deep communities. This proves the great capacity of adaptation of microeukaryotes to different environments. In addition, the different impact of the culturing bias in surface and deep communities remains as an intriguing observation worth to be further addressed.

Ubiquity: Everything is everywhere?

Finlay *et al*. (2004) stated the difference between "ubiquity" and "ubiquitous dispersal". They claimed that most protists have ubiquitous dispersal, implying that they are not necessarily found everywhere but should be present in suitable habitats around the world (Caron *et al.* 2009). Another important aspect of this debate is that the presence of the same sequence in separate oceans can give us information about the ubiquity of the respective species whereas the absence can be simply due to undersampling or to temporal successions. In the Malaspina dataset, 42 OTUs are present in the total of 27 stations, and these represent 80% of the pyrosequences. This qualitative view of communities is pretty close to the idea that "everything is everywhere". However, following "the environment select" concept, the distribution of these OTUs is not uniform, and an OTU that belongs to the rare biosphere in one sample often is the dominant OTUs in another, as observed in several examples, for instance with Fungi. As commented in the introduction, the environmental homogeneity of the deep ocean is an old and misleading idea and, despite putative high dispersal ability, there is a strong environmental effect on community composition. Probably the intrinsic properties of deep water masses plus the presence of preys or alternative food allows

the existence of different trophic niches.

Phagotrophy: the relation with prokaryotes

Phagotrophy was expected to be the principal trophic pathway for microeukaryotes in the deep ocean. To test this hypothesis, we first analyze the relation between the abundance of prokaryotes and microeukaryotes, and second the relation between their ratio and the diversity. Considering the entire dark water column, the abundance of microeukaryotes correlates well with that of prokaryotes (R^2=0.50, p=0.0001), except in South Pacific stations (R^2=0.08, p>0.05, Figure 5, Chapter 3). Subsequent multiple regression analyses proved that this result is independent from depth (Chapter 3). However, despite this significant relationship, prokaryotes explain only a part of the variance of the microeukaryotes abundance. We use prokaryotes abundance also to explain the variability in diversity but the result is not significant (p>0.058), although 34% of the variability is significatively explained by the ratio Prokaryotes to microeukaryotes (p=0.001). Low values of this ratio, similar to the ones in epipelagic waters (ca. 2000), correspond to communities dominated by Collodaria and Chrysophyceae, suggesting a putative grazing role for these two classes. Considering that less than half of the variability of abundance and diversity is explained by prokaryotes and that the abundance ratio has an average value higher than at surface, we conclude that phagotrophy in the deep ocean seems to give also a significant space for other trophic pathways such as osmotrophy and parasitism.

Osmotrophy: the role of Fungi

Comparing the prokaryotes to microeukaryotes abundance ratio with the relative abundance of several taxa, only Fungi presents a significant relationship (Figure 7, Chapter 3). Considering that Fungi are proved osmotrophs and probably unable to perform phagotrophy (Richards *et al*. 2012), this suggests that where the community is dominated by Fungi there is a lower grazing pressure on prokaryotes, and this could favour the high abundance ratio.

As seen in the introduction (Figure 7b) the distribution of DOC is not constant in the bathypelagic region. In fact, DOC decreases along the deep conveyor belt resulting in an overall higher concentration in Atlantic than in Pacific waters. It is possible to associate part of the decrease of DOC from Southern Ocean to North Pacific to the presence of Fungi in these waters. However, considering that DOC is more concentrated in the Atlantic Ocean, especially in the north, is difficult to understand why Fungi generally do not thrive in these waters. I present next several possible explanations for the absence of Fungi in the Atlantic Ocean:

- *Fungi versus prokaryotes*. A first simple hypothesis is that prokaryotes are the main con-

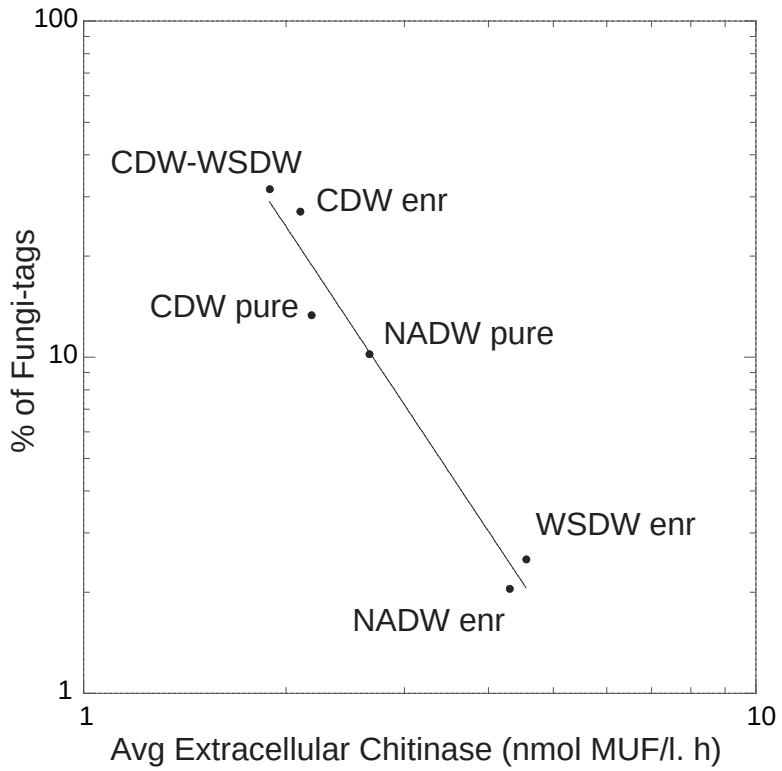


Figure 1. Relationship between the average extracellular chitinase concentration and the average fungal tag abundance in the different water masses defined.

sumers of DOC in Atlantic waters and outcompete Fungi. An antagonistic relation of Fungi and bacteria has been seen in previous experiments, as for example in Moller *et al.* (1999) where Fungi and Prokaryotes clearly compete for the same DOC.

- *Fungi versus Chrysophytes*. Chrysophytes can be phagotrophs and osmotrophs (Sandgren *et al.* 1995, Sanders *et al.* 2001), and they may survive thanks to a combination of these two strategies. Except in station 32 they never share dominance with Fungi. Våge *et al*. (2013) built a model to test the importance of mixotrophy compared with pure osmotrophy. They demonstrated that at low size ratio between prey (prokaryotes) and predators (Chrysophytes), as occurs in the deep ocean, for a mixotroph is very convenient the "eating the competitor" strategy (Thingstad *et al.* 1996). So in this case mixotrophs (Chrysophytes) could suppress pure osmotrophs (Fungi) by foraging on them, and this could happen in Atlantic and North Pacific waters.

- *Fungi and recalcitrant carbon*. The presence of Fungi in older waters poorer in DOC could be explained by their specialization to assimilate recalcitrant carbon. A possible mechanism is the secretion of superoxide molecules, in particular oxided forms of Mn, which oxidizes recalcitrant carbon to more bioavailable forms (Hansel *et al.* 2012). This mechanism is probably shared with bacteria. So, another explanation of higher ratio of

Prokaryotes to microeukaryotes is that in a Fungi dominated community prokaryotes could also thrive on recalcitrant DOC.

- *Fungi versus chitinase*. One of the typical features of Fungi is the presence of chitin in the outer wall (Richard *et al*. 2012). During the Malaspina expedition we measured extracellular enzymatic activities, including chitinase, the enzyme that digests chitin. The chitinase activity turns out to be higher in the Atlantic than in Pacific and Indian Oceans. Considering averaged vaules for each defined watermass, the relative abundance of Fungi tags show an inverse relationship with the chitinase activity (Figure 1) in a highly significant correlation ($p=0.0005$; R^2 of 0.95). Chitinase enzymes could be produced by prokaryotes as well as microeukaryotes (Cottrell *et al*. 2000). Although it is unknown if Fungi are the principal targets of this enzyme, the extracellular chitinase probably produces an environment not favourable for their life.

Osmotrophy is also present in other taxonomic groups, such as the labyrinthulids (Raghukamar *et al*. 2001) or Excavata (Lara *et al.* 2009). Indeed, the extent of the osmotrophic process should be studied to understand the impact of heterotrophic protists in the global carbon balance.

Parasitism: the hidden relationships

The inference of parasitic interactions from sequencing data is quite hard because there is no clear correspondence between host and parasite abundance (Skovgard *et al.* 2014). And, more severely, a naked sequence sometimes tells very little about ecological performance. Several marine microeukaryotic clades are considered to be parasites, being MALV-I and MALV-II the most abundant. Several species within dinoflagellates and Fungi can also be parasites. The supposed hosts of these parasites are other microeukaryotes and also macrofauna. In our dataset, the relative abundance of MALV-II tags have a significant correlation with the relative abundance of Metazoan tags ($R^2=0.45$, $p=0.0005$) (Figure 2). This relation is better for MALV-I ($R^2=0.60$) and less strong for dinoflagellates ($R^2=0.41$), but undetected for fungi or the other classes. The three relations are particularly evident in Atlantic samples, where R^2 is respectively 0.85, 0.89 and 0.76. Generally parasitism is strongly present in Fungi and probably also in its marine clades, especially since the principal OTU is highly similar to parasitic species. Indeed, *Engyodontium album* is *a* parasite of the *Felis domesticus* (Dennis 1995). Nevertheless, assuming a random distribution of the hosts it is expected a random distribution of parasite tags. In this sense, the distribution of Alveolate tags better fits with this scenario compared with the non-random Fungi distribution.

The last decades of protist research have focused on the study of the diversity of microeukaryotes,

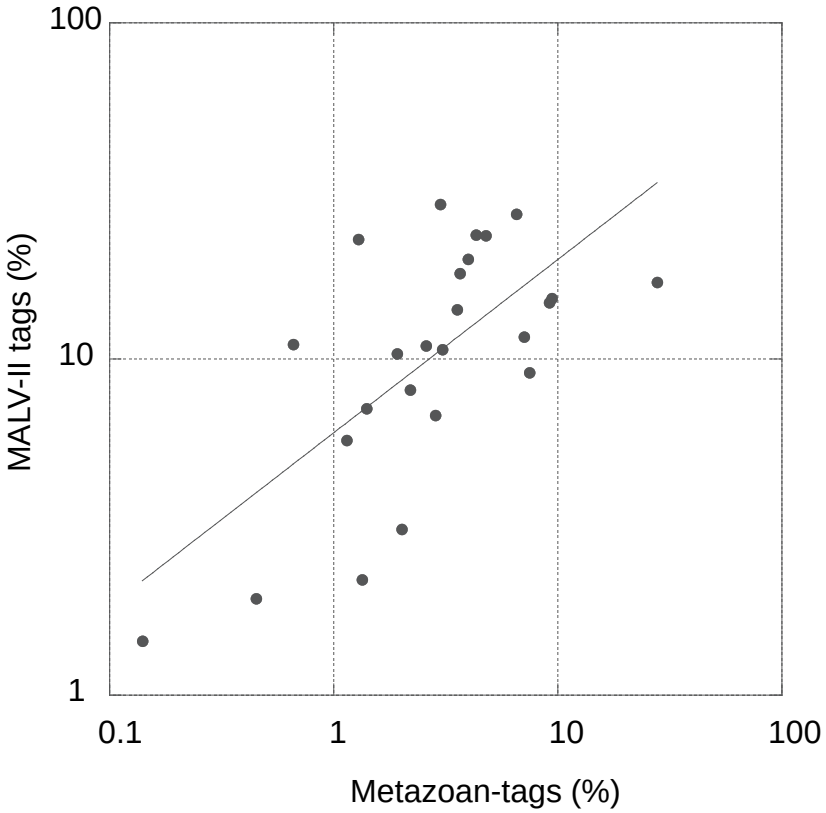


Figure 2. Relationship between the relative tag abundance of MALV-II against that of metazoans.

at the beginning defined as "unexpected" in terms of high and novel diversity. Now that we expect such amazing complexity of taxa composition, the attention is moving to the investigation of the function of the different taxa in the ecosystem. The assignation of a clear ecological role, through the classical method of culturing and new single-cell genomic approaches, will soon improve the vision that we have now on such important environmental players.

Conclusions

1) The V4 region of the 18S rDNA is better than the V9 region for representing the variability of the entire gene. On average, the variability contained in the V4 region is 1.4 times that of the complete 18S rDNA gene.

2) Typically, the maximal genetic distance among sequences belonging to a same eukaryotic taxonomic class is 0.25. This value can be used to assess the taxonomic equivalence of environmental ribogroups.

3) Epipelagic microeukaryotes communities are typically formed by Alveolata (47% of sequences), Stramenopiles (19%), and Rhizaria (13%). Additional groups belong to CCTH and Archaeplastida. Often Fungi and Excavata are very low (less than 1%) or undetected.

4) The abundance of microeukaryotes averages 54 cells mL^{-1} in mesopelagic samples and 14 cells mL^{-1} in bathypelagic samples. This variability is explained mainly by depth, prokaryotes abundance and oxygen concentration.

5) The cell size of planktonic microeukaryotes increases with depth. Cells larger than 4 μm represent 12% at 200 m and 22% at 4000 m. Total biomass ranges from 280 pg C mL^{-1} in the upper mesopelagic layer to 50 pg C mL^{-1} in the deepest layer.

6) The diversity assessed by 454 pyrosequencing compares very well with a parallel metagenomic survey. In general, the percentage of eukaryotic supergroups is very similar by the two approaches (Figure 7a) and the same OTUs are retrieved. The main difference is a much higher representation of Excavata in miTags and a lower representation of Alveolata.

7) Four abundant classes generally compose the bathypelagic community of microeukaryotes: Collodaria, Chrysophyceae, MALV-II and Basidiomycota. However, the relative abundance of these classes varies a lot among samples.

8) The variability in community composition between samples is well explained by the water mass they belong (26% of variability) and by the abundance ratio between prokaryotes and microeukaryotes in the respective samples (34%).

Resumen de la tesis

Una breve historia sobre la investigación de los protistas marinos

La definición de "protista" abarca una gran diversidad de organismos. En términos generales, los protistas son microorganismos eucarióticos. De hecho, esta definición no tiene un verdadero significado evolutivo ya que incluye todos los eucariotas que no son animales, plantas u hongos. La primera observación de protistas registrada fue la de Leeuwenhoek en el año 1674, pero el termino fue acuñado y popularizado en 1866 por Haeckel (famoso por sus ilustraciones detalladas de estos organismos, Figura 1). Al principio este termino comprendía también los organismos procariotas.

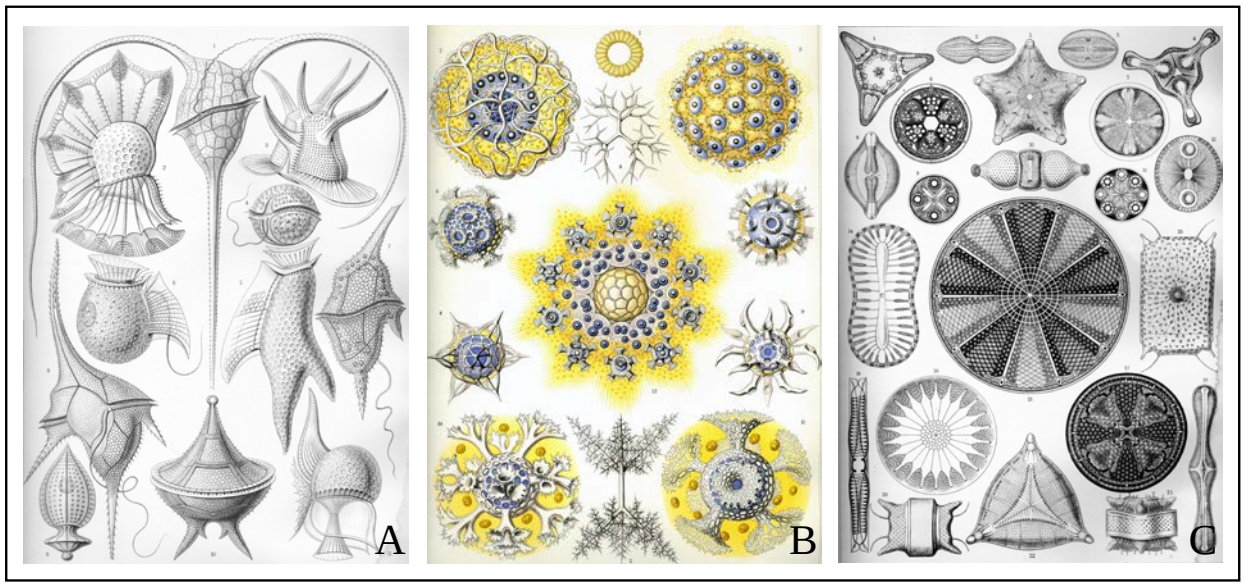


Figura 1. La diversidad morfológica en Alveolata (A), Rhizaria (B) y Stramenopiles (C). Dibujos de Ernst Haeckel, 1904.

Protista fue entonces considerado como un nuevo reino, de la misma manera que *Animalia* y *Plantae,* aparentemente menos importante y con menos categorías. Hoy en día, gracias a numerosos estudios que empezaron en la segunda mitad del siglo XX, sabemos que animales y plantas son dos pequeñas hojas en el árbol filogenético de los eucariotas, que está compuesto sobretodo por formas de vida unicelulares (Figura 2).

El progreso tecnológico ha hecho posible el paso del estudio de lo visible hacia el descubrimiento

de lo invisible. La investigación de los microorganismos se fundamenta en diferentes técnicas. Un método clásico es la observación y el recuento de células de protista con microscopio de epifluorescencia, que implica la utilización de marcadores celulares, como por ejemplo el DAPI (4',6-diamino-2-fenilindol) que se une al ADN de las células (Porter y Feig 1980). Una técnica mas precisa es la Hibridación Fluorescente In Situ (FISH, Pernthaler *et al*. 2002, Massana *et al*. 2006). Esta técnica, que gracias a sondas de oligonucleótidos taxón-específicas detecta sólo determinadas células, permite reunir informaciones acerca de la abundancia y la diversidad global de los protistas marinos (Morgan-Smith *et al*. 2011 y 2013), y puede ser útil también en experimentos de depredación (Fu *et al*. 2003, Jezbera *et al*. 2005, Massana *et al*. 2009). Otro método para cuanti-

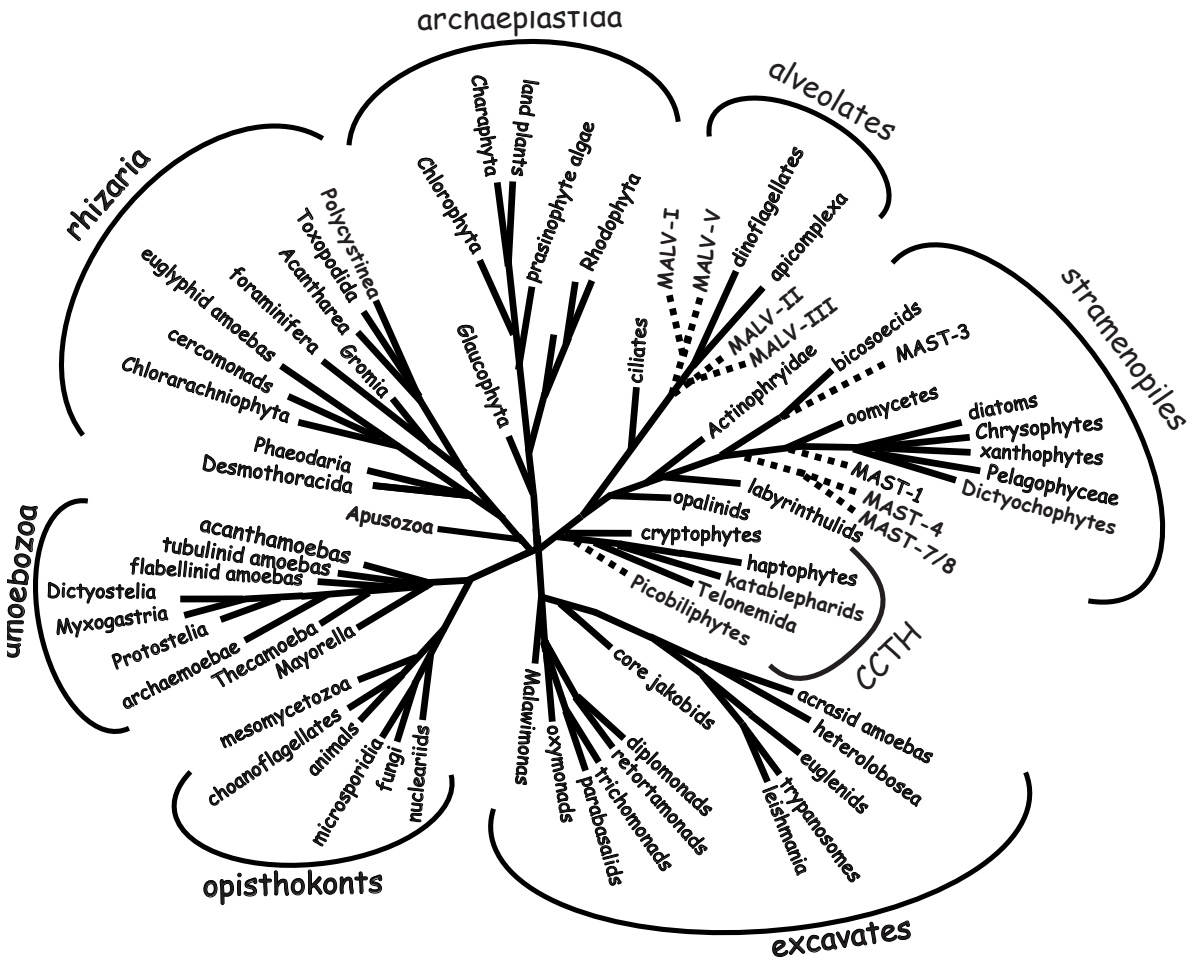


Figura 2. Árbol de la vida de los eucariotas, que muestra la filogenia de los principales grupos de eucariotas basados en datos moleculares y ultraestructurales (adaptado de Baldauf 2003). Las líneas punteadas indican las posiciones de los principales linajes conocidos gracias a estudios moleculares independientes de cultivos.

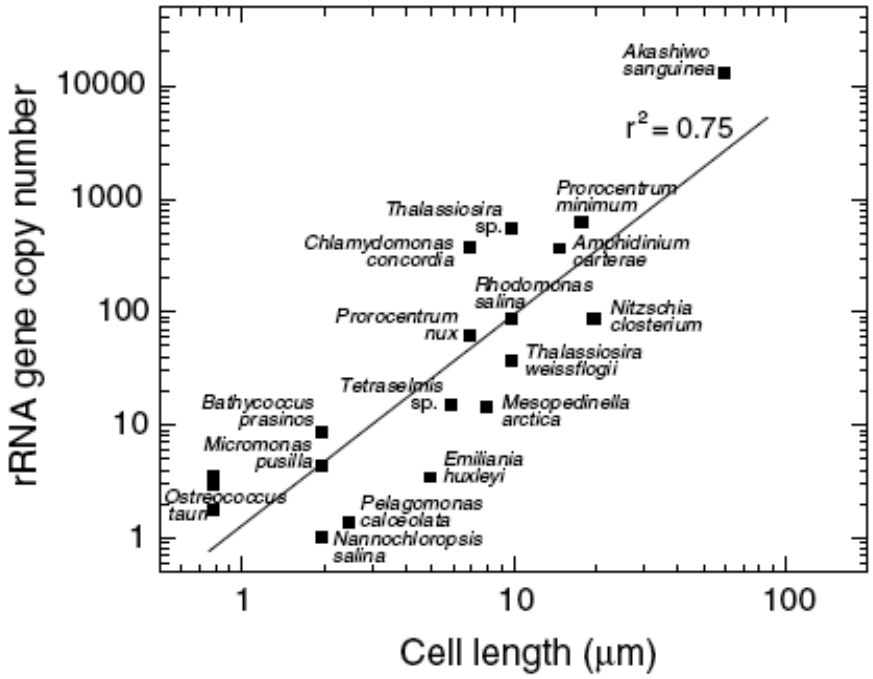


Figura 3. Correlación entre el tamaño de la célula y el número de copias de ADNr en diferentes especies de protistas. Tomado de Zhu et al. (2005).

ficar la abundancia microbiana es la citometría de flujo (Zubkov *et al*. 2006 and 2007; Christaki *et al*., 2011). Esta técnica es muy útil en caso de numerosas muestras, pero aún necesita ser mejorada y optimizada por algún grupo funcional como el de los microeucariotas heterotróficos. A pesar de la posibilidad de identificar alguna topología celular, la microscopia de epifluorescencia no es bastante precisa para el estudio de la diversidad. De hecho, los caracteres morfológicos son inútiles para identificar células de tamaño *pico-*, también para la mayoría de tamaño *nano-* y a menudo no permite investigar en profundidad la taxonomía a nivel de especie.

Con el desarrollo de los métodos moleculares el estudio de la diversidad microbiana aumentó exponencialmente. Estudios pioneros en esta dirección fueron los de Díez *et al.*, 2001, López-García *et al.*, 2001 y Moon-van der Staay *et al.*, 2001. Estas investigaciones tenían que enfrentarse todas con un único problema general: para estimar la diversidad es fundamental identificar grupos de

Reference	Species	Sim	Origin
Rooney *et al.* (2004)	*Cryptosporidium parvum*	92.1	Genome
	Plasmodium Falciparum	89.5	Genome
	Plasmodium berghei	92.0	GeneBank
Alverson *et al.* (2005)	*Skeletonema grethae*	99.2	Strain
	Skeletonema japonicum	99.4	Strain
	Skeletonema menzeleii	99.4	Strain
	Skeletonema pseudocostatum	99.5	Strain
	Skeletonema subsalsum	99.5	Strain
Simon *et al.* (2008)	*Phoma exigua*	99.5	Strain
	Mycospharella punctiformis	99.6	Strain
	Teratospheria microspora	99.6	Strain
	Davidiella tassiana	99.4	Strain
	Aspergillus nidulans	99.6	Strain
Gong *et al.* (2013)	*Tintinnopsis sp.*	99.1	Individual
	Pseudotontonia sp.	99.3	Individual
	Strombidium sp.	99.7	Individual
	Vorticella sp.	99.1	Individual

Tabla 1. Variabilidad intragenomica del gen 18S ADNr (SSU). Se muestran los valores de similitud genética intragenomica en diferentes especies de microeukaryotes. Los valores bajos de similitud en *Plasmodium spp.* Y *Cryptosporidium parvum* se explica por la presencia efectiva de las diferentes formas ribosomales que se activan en diferentes huéspedes de estos parásitos.

organismos similares definidos como “especies” en la taxonomía clásica. Diferentes conceptos de especie han sido aplicados a los microorganismos en general y a los protistas en particular (Roselló-Mora and Amann 2001, Schlegel and Meisterefeld 2003). La definición mas pragmática propone que una especie es “un grupo de organismos que comparten características morfológicas similares”. La definición biológica, quizás la más útil para animales y plantas, define una especie como “un grupo de organismos capaces de cruzarse sexualmente y producir una progenie fértil”. Aunque la división celular asexual es la mas común entre los protistas, también hay diferentes casos de reproducción sexual (Amato *et al.*, 2007), pero existe poca información acerca de cuán distribuida se encuentre entre los diferentes clados y con qué frecuencia ocurre. Por supuesto, estudios acerca del ciclo de vida de las especies de protistas para encontrar la incidencia de la reproducción asexual o sexual son muy necesarios. La principal limitación de este tipo de estudios es que relativamente pocas especies de protistas están cultivadas y bien caracterizadas e incluso

algunos grupos carecen por completo de representantes cultivados. Por lo tanto, el uso del concepto biológico de especie no es muy practico para estudiar la diversidad de los protistas.

Afortunadamente para los microbiólogos, en los años 70 del siglo pasado Carl Woese se dio cuenta de la posibilidad de identificar todas las formas de vida comparando sus secuencias de ADN (Woese *et al*., 1977). Esta operación se compone básicamente de dos pasos: el alineado de las secuencias de ADN del mismo gen y la medida de sus distancias genéticas. La diana favorita de esta operación desde el principio fue el gen ribosomal del ARN (rDNA). Este gen está presente en todos los organismos y está suficientemente conservado para ser usado en una filogenia entre cualquier forma de vida. La taxonomía molecular presenta varias ventajas: se puede aplicar a una amplia gama de taxones, a todas las etapas de la vida y a grandes volúmenes de datos que son típicos de la mayoría de los estudios ecológicos (Caron *et al*. 2009). El rADN es muy útil pero no es un blanco perfecto, considerando que es un gen multicopia, particularmente en los eucariotas. En cepas de algas el número de copias varía de 1 a 10.000 (Zhu *et al*. 2005), lo que implica que la abundancia relativa de genes puede desviarse considerablemente de la abundancia relativa de células. El número de copias es proporcional al tamaño celular y del genoma (Figura 3), por lo tanto la posibilidad de grandes variaciones es más baja para las células de tamaño pico- y nano-. Además, es posible que estas copias tengan una gran variabilidad a nivel intragenómico. El riesgo de la variabilidad intragenómica es que se podrían detectar dos o más secuencias diferentes donde en realidad sólo hay un organismo. Una vez más, en la mayoría de los casos, esta variabilidad intragenómica es muy baja (Tabla 1).

La importante novedad de las técnicas moleculares fue la posibilidad de un estudio más realista de la diversidad microbiana marina, particularmente a nivel del nano- y pico plancton. El método clásico fue la construcción de bibliotecas de clones del gen 18S rDNA, las secuencias se amplificaban

a partir de ADN genómico ambiental, con una reacción en cadena de la polimerasa (PCR, Mullis *et al*. 1983). Típicamente, desde una biblioteca de clones se conseguían entre 100 y 500 secuencias. Estas secuencias ribosomales se convirtieron en la base de una nueva taxonomía molecular; de hecho fue creado un nuevo concepto de especie más pragmático que el biológico o morfológico: la unidad taxonómica operativa (OTU). Siguiendo este criterio, las secuencias fueron agrupadas en unidades contables que tenían un cierto criterio de divergencia genética, elegida por el investigador, en una operación normalmente llamada "clustering" (agrupación). Es importante subrayar que la manera de agrupar las secuencias en OTU es un paso crucial para determinar nuestra visión de la diversidad en muestras marinas.

A pesar de alguna limitación, el gen rDNA (y particularmente el 18S rDNA) sigue siendo el mejor compromiso para estudiar la diversidad de protistas y ha sido elegido como diana en la emergente tecnología de secuenciación de alto rendimiento (*high-throughput sequencing*, HTS), 454 e Illumina, que ha evolucionado tan rápidamente que la definición de "secuenciación de próxima generación" ha quedado obsoleta en menos de 5 años. El número de secuencias recogidas con las HTS es de varios órdenes de magnitud mayor que las conseguidas mediante bibliotecas de clones, lo que conlleva a la aparición de nuevos problemas relacionados con la gestión de este gran número de datos. Sin embargo, hay un gran entusiasmo relacionado con la posibilidad de poder secuenciar el océano (Venter *et al*. 2004) y muchos científicos están trabajando en la optimización del método para aumentar la confianza en él (Kunin *et al*. 2009, Quince *et al*. 2009). Las nuevas tecnologías de secuenciación dan la posibilidad de estudiar la diversidad de manera más profunda. Es importante recordar que, cuando están combinadas con la PCR, las HTS sufren los mismos tipos de errores que las bibliotecas de clones (Winzingerode *et al*. 1997). Hoy en día la metagenómica, aunque tiene por objetivo principal más el estudio de las funciones metabólicas que el de la diversidad de las

especies, es una posible alternativa para recolectar secuencias de 18S rDNA desde comunidades naturales de microorganismos (Logares *et al*. 2013). El uso de las técnicas metagenómicas es independiente del paso por la técnica de PCR, eliminando así esta fuente de errores. Para un estudio profundizado de las características de la diversidad de los protistas, todos los métodos descritos han sido utilizados en esta tesis.

Taxonomía general de los microeucariotas

La taxonomía de los eucariotas es un continuo motivo de debate en el mundo científico (Burki *et al*. 2008). Gracias al uso conjunto de microscopía y biología molecular es posible identificar la mayoría de los taxones, pero el verdadero reto es entender cómo estos grupos se relacionan entre sí. En este trabajo hay una mezcla de grupos morfológicos clásicos (mejorando su definición gracias a las herramientas moleculares) y nuevos ribogrupos, que están formados por secuencias que se agrupan juntas en un árbol filogenético fuera de los grupos conocidos. Estos nuevos ribogrupos se insertan entre medio de los grupos clásicos, que son los que representan la espina dorsal del árbol de la vida eucariota y están definidos por su morfología. La taxonomía general usada por la mayoría de los grupos morfológicos está basada en Adl *et al*. (2012).

De todo el árbol de la vida (Figura 2) cuatro supergrupos de protistas merecen ser mencionados debido a su particular importancia en los estudios moleculares ambientales: Alveolata, Rhizaria, Stramenopiles y CCTH. Alveolata, a menudo el supergrupo más abundante, comprende dos de los clados más clásicos: dinoflagelados y ciliados. Rhizaria está compuesto sobretodo por radiolarios, que pueden ser solitarios o vivir en colonias y se caracterizan por una estructura compleja, Cercozoa (también conocidos como Filosa) y Foraminífera, un grupo que prefiere los sedimentos. Los estramenópilos comprenden grupos fototróficos como las diatomeas y grupos heterotróficos como los bicosoécidos. El CCTH es un grupo propuesto recientemente que incluye Cryptophyta, Centroheliozoa, Telonemia y Haptophyta (Burki *et al*. 2009), sin embargo, estudios filogenéticos mas recientes levantan dudas sobre su monofilia (Baurian *et al*. 2010, Burki *et al*. 2012). En las últimas décadas, y gracias a las investigaciones moleculares, "un grupo" ha recobrado importancia en el ecosistema oceánico, los hongos (Bass *et al*. 2007, Richards *et al*. 2012). Tradicionalmente, los Fungi han sido estudiado más por botánicos que por protistólogos. Sin embargo, considerando que

muchas especies de hongos son unicelulares, entran perfectamente en la definición de microeucariotas, objetivo de nuestro estudio. Los Fungi han sido encontrados en la columna de agua y en los sedimentos (Bass *et al*. 2007, Lepere *et al*. 2008 Jebaraj *et al*. 2009, Edgcomb *et al*. 2011, Richards *et al*. 2012). Inicialmente una práctica común era eliminar las secuencias de Fungi de las bibliotecas de clones de los eucariotas, como normalmente se hace con los Metazoa. Ahora estas secuencias son apreciadas y guardadas, de hecho no tendría sentido excluir este importante elemento del ecosistema marino.

Los ribogrupos representan una gran parte de las secuencias encontradas en los estudios moleculares. De hecho la mayoría de las secuencias de este estudio pertenecen a los alveolados marinos (MALV), que fueron ya encontrados en el primer estudio molecular del océano profundo (Lopez-Garcia *et al*. 2001) y mejorada su definición posteriormente (Groissillier *et al*. 2006). Otros ribogrupos importantes son los estramenópilos marinos (MAST), definidos en el año 2004 por Massana *et al*., y los Picozoa (conocidos anteriormente como Picobiliphyta) que fueron identificados primero por sus secuencias ambientales (Not *et al*. 2007) y posteriormente cultivados (Seenivasan *et al*. 2013). Los Rhizaria abarcan tres ribogrupos, RAD_A, RAD-B, y RAD_C, el segundo comprende el grupo morfológico de Sticholonche conocido anteriormente como Taxopodia. Todos estos ribogrupos se encuentran frecuentemente y están ampliamente reconocidos, entrando de lleno en la taxonomia “práctica”.

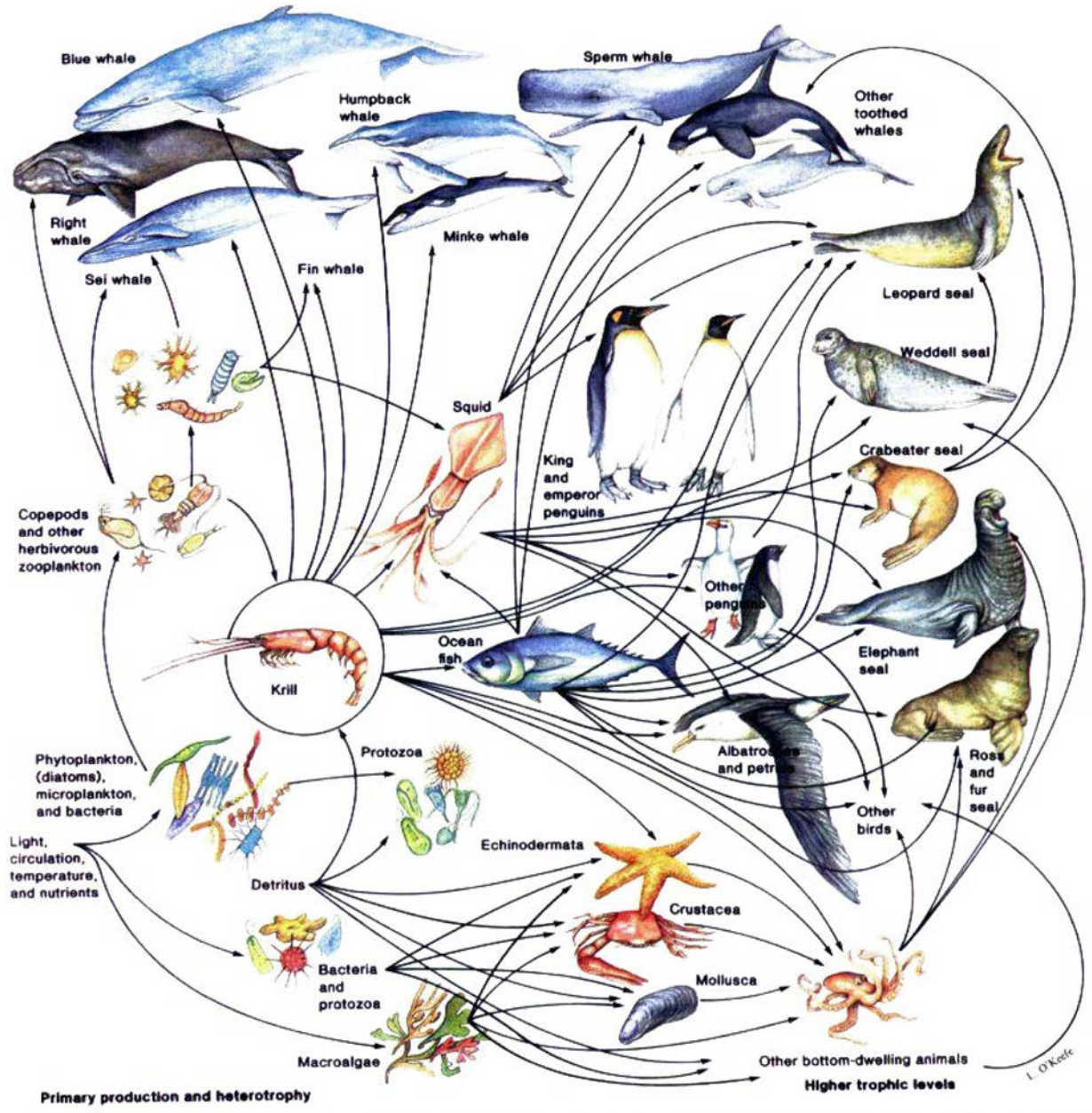


Figura 4. Una red trófica marina completa, lo que indica una gran variedad de especies y sus interacciones. La red trófica microbiana se muestra a la izquierda. Dibujo de O'Keefe L.

Papel trófico y participación en los ciclos biogeoquímicos

En un mililitro de agua epipelágica hay entre 1000 y 10.000 células de microeucariotas. Es difícil identificar la función ecológica de cada taxón pero está claro que juntos juegan un papel importante en los ciclos biogeoquímicos, tanto como autótrofos o como heterótrofos. Es importante recordar que el fitoplancton, hoy en día considerado como la suma de las bacterias fototróficas y

de los protistas fototróficos, produce el 70% del oxígeno total del planeta (Epstein *et al*. 1993), haciendo posible la vida en la tierra. Generalmente los organismos unicelulares están conectados en una compleja red trófica basada en el tamaño. El bucle microbiano, propuesto por Azam *et al*. en 1983, constituye un interesante indicio de esta red. Gracias a este bucle, la materia orgánica disuelta es consumida por los procariotas y llega a los niveles tróficos superiores debido a que los protistas fagotróficos se alimentan de procariotas y son a la vez presa de organismos zooplanctónicos más grandes (Figura 4 y Figura 5b). Los procariotas se pueden considerar como la máquina bioquímica que guía los principales ciclos biogeoquímicos (carbón, nitrógeno, azufre), pero lo que controla finalmente la velocidad de estas reacciones metabólicas son los protistas bacterívoros, probablemente junto a los virus (Boras *et al*. 2010).

La fagotrofía, la ingestión de partículas de comida mediante invaginación de la membrana celular, está muy difundida en los taxones de protistas. Es un proceso que está bastante bien estudiado tanto en el ambiente (sobretodo en agua epipelágica) como en los cultivos. Las clases importantes de fagotrofos son por ejemplo los ciliados, los cuales representan a los predadores en la clásica red trófica, las crisófitas y el ribogrupo MAST-4, quizás el bacterívoro mas importante en el bucle microbiano marino (Massana 2011). Sin embargo, la fagotrofia no es la única forma de heterotrofia en el ambiente. Diferentes especies que pertenecen a los Fungi (Richards *et al*. 2012), Excavata (von Der Heyden *et al*. 2004), Chrysophyceae (Sandgren *et al.* 1995, Sanders *et al.* 2001) o Labyrinthulidae (Raghukamar *et al*. 2001) sobreviven gracias a la osmotrofia, que es la asimilación de compuestos orgánicos disueltos mediante ósmosis. Además, hay varios ejemplos de grupos que sobreviven en el océano gracias a interacciones parasitarias y abarcan un ancho abanico de organismos anfitriones. Estos parásitos marinos incluyen el ribotipo MALV-I y -II , que es el grupo más grande en término de secuencias encontradas (Siano *et al*. 2010). El estudio ambiental

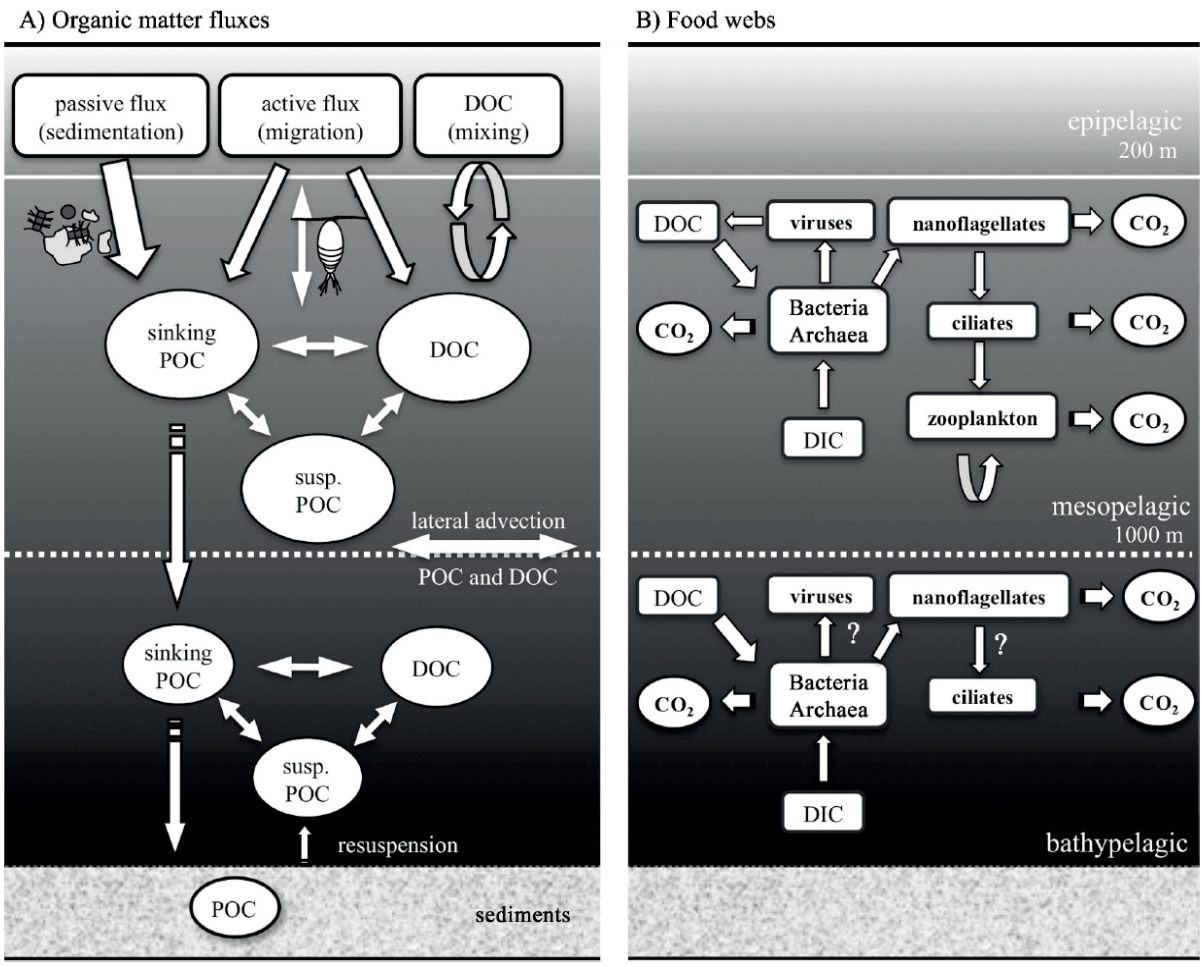


Figura 5. Representación esquemática de los flujos de materia orgánica (A) y la red trófica microbiana (B) en el ecosistema oceánico profundo (de Aristegui et al. 2009). A) Se indican tres conjuntos de carbono orgánico interconectados: carbono orgánico disuelto (DOC), carbono orgánico particulado que se hunde (POC) y POC suspendidos. B) Red trófica microbiana de los reinos mesopelágicos y batipelágicos. Los procariotas en el océano oscuro pueden sobrevivir gracias al DOC (heterotrofía) y carbono inorgánico (quimiosíntesis). En la zona batipelágica el control de los procariotas por los flagelados o los virus y el papel de los ciliados sigue siendo enigmático (signos de interrogación).

del parasitismo es bastante difícil y a pesar de los pocos casos documentados (Chambouvet *et al*. 2008), la magnitud de este fenómeno, como por ejemplo la osmotrofia, no es clara. Fuera de estas estrictas diferencias tróficas, hay que recordar que el mundo unicelular favorece una cierta plasticidad en el método trófico, y la mixotrofia, la combinación de diferentes estilos, parece ser una conducta común en los grupos de protistas (Sanders *et al.* 1991; Jones 2000, Zubkov *et al*. 2008).

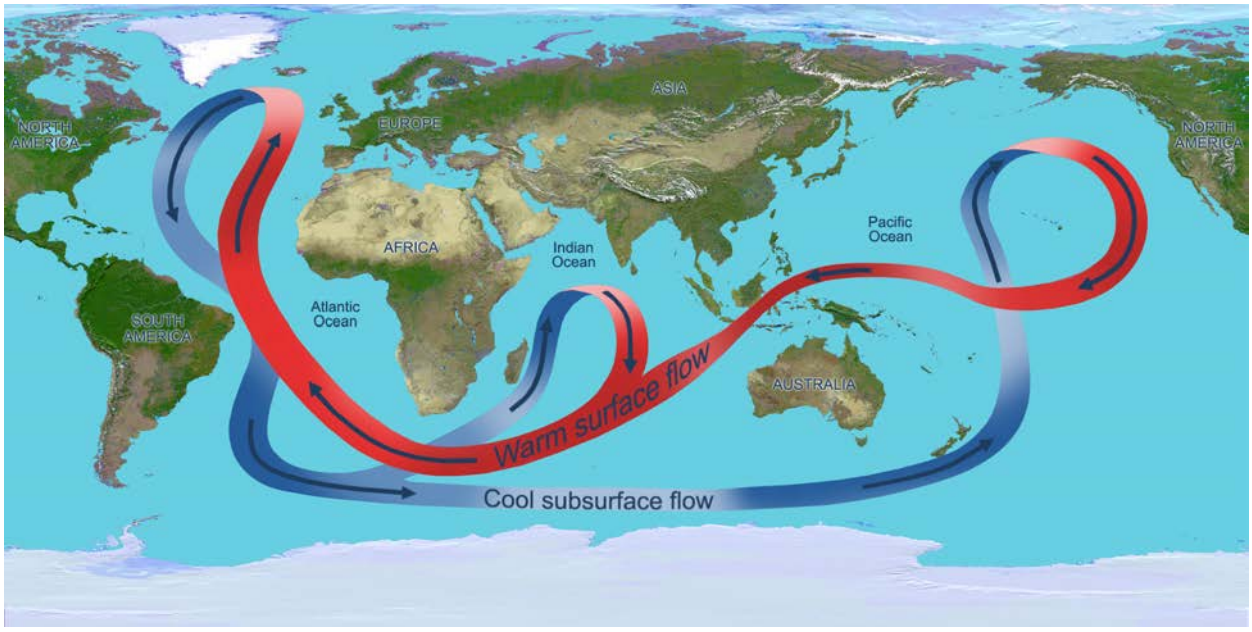


Figura 6. La circulación termohalina. Masas de aguas frías y densas se hunden en el Atlántico Norte y en los océanos del sur, creando una corriente que fluye en las cuencas oceánicas. Estas aguas vuelven a la superficie gracias al influjo de los afloramientos en los océanos Índico y Pacífico Norte, formando una corriente de agua cálida que fluye en la dirección opuesta a las capas superiores. Cerca del Polo Norte las aguas se enfrían y se hunden volviendo a reiniciar un ciclo que dura siglos.

El océano profundo: un hábitat peculiar

La capacidad de adaptación de los microeucariotas es evidente con el hecho de que están presentes en todo el planeta, incluyendo todos los tipos de ambientes extremos. En el océano, sabemos que están presentes en toda la columna de agua (Not *et al*. 2007). Sin embargo, por razones obvias, las comunidades superficiales se estudian más que las profundas. De hecho, el funcionamiento del océano profundo todavía está alejado de esclarecerse completamente. Tradicionalmente el océano profundo esta dividido en tres zonas, la mesopelágica (200-1000 m), la batipelágica (1000-4000 m) y la abisopelágica (mas de 4000 m). La región mesopelágica, donde a menudo se encuentra la termoclina, está más influenciada por los aportes epipelágicos (0-200 m) que los dos estratos más profundos. De hecho, una gran parte del carbono orgánico fijado por la fotosíntesis es respirado en esta zona (Aristegui *et al*. 2005).

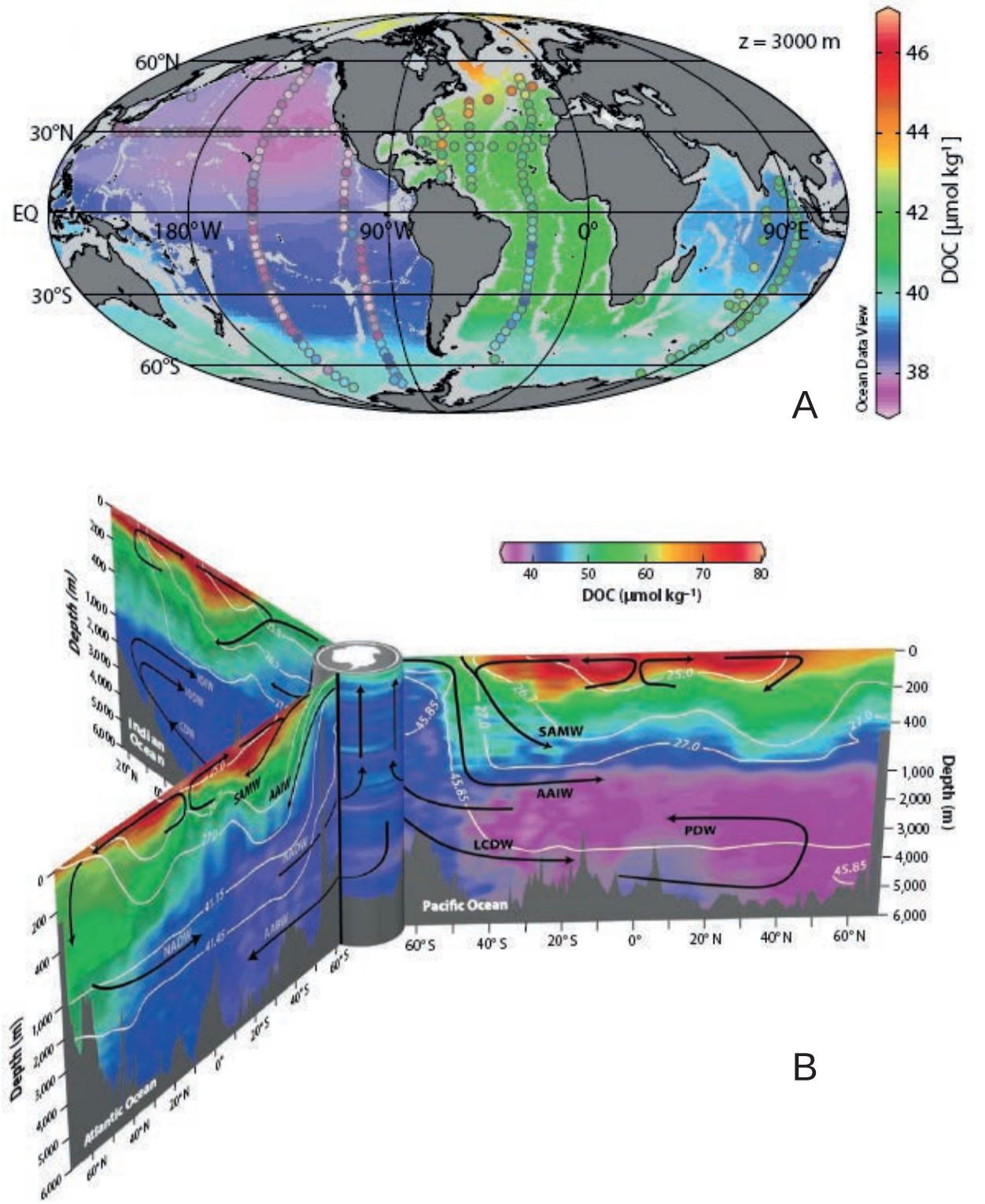


Figura 7. Distribución de carbono orgánico disuelto (COD μmol kg-1) en los océanos del mundo (Hansell et al 2009). A) Distribución del COD a 3000 m. Los puntos son los valores observados, mientras que los colores del fondo provienen de un modelo. B) Distribución del COD en el Atlántico central, Pacífico central y en el Índico oriental. Las flechas representan la circulación de masas de agua.

La zona batipelágica muestra varias diferencias en comparación con los ecosistemas superiores. Teniendo en cuenta los parámetros físicos, este sistema es más estable: el agua está generalmente bien oxigenada (aunque existan zonas anóxicas), la temperatura presenta un rango muy estrecho a nivel mundial, de 1 a 4 ºC, y la salinidad es prácticamente constante alrededor de 35 ppm. En la región batipelágica, la presión es muy alta (5 a 10 MPa), pero esto no limita el desarrollo de la vida a escala macro y micro. A pesar de esta aparente homogeneidad, todavía es posible reconocer diferentes masas de agua en base a los parámetros físicos y químicos, siendo las más importantes la del agua profunda del Atlántico Norte (NADW), la del agua profunda circumpolar (CDW) y la del agua profunda del mar de Weddel (WSDW). Estos tres tipos de agua se encuentran respectivamente en el Atlántico, Pacífico e Índico.

Desde un punto de vista químico, la concentración de la materia orgánica, de los nutrientes inorgánicos y de otros compuestos químicos puede ser muy variada en diferentes regiones marinas, dependiendo del material que se hunde desde la superficie. Generalmente, el océano batipelágico es rico en las formas oxidadas de nutrientes inorgánicos (NO_3, PO_4) y carece de compuestos reducidos tales como el amonio (Nagata *et al* . 2010). A nivel mundial las profundidades del océano son consideradas como la mayor reserva de carbono orgánico biodisponible (Libes 1992, Benner 2002), sin embargo, la concentración de carbono orgánico disuelto (COD) difiere entre las diferentes cuencas (Hansell y Carlson 1998). El papel de las profundidades del océano como sumidero de carbono inorgánico es bastante intuitivo (Figura 5a). Entre el 5 y el 15% del carbono fijado por la fotosíntesis en capas marinas superiores llegan al reino batipelágico a través de la bomba biológica (Giering *et al*. 2014), donde es respirado y secuestrado durante siglos hasta que regresa a la parte superior del océano y luego al ambiente. Por lo tanto, el sistema batipelágico tiene un papel muy importante en el balance global de CO_2 y, teniendo en cuenta su vínculo con problemas críticos,

tales como el calentamiento global y el cambio climático, es muy importante definir el destino final del COD del agua profunda (Aristegui *et al*. 2009).

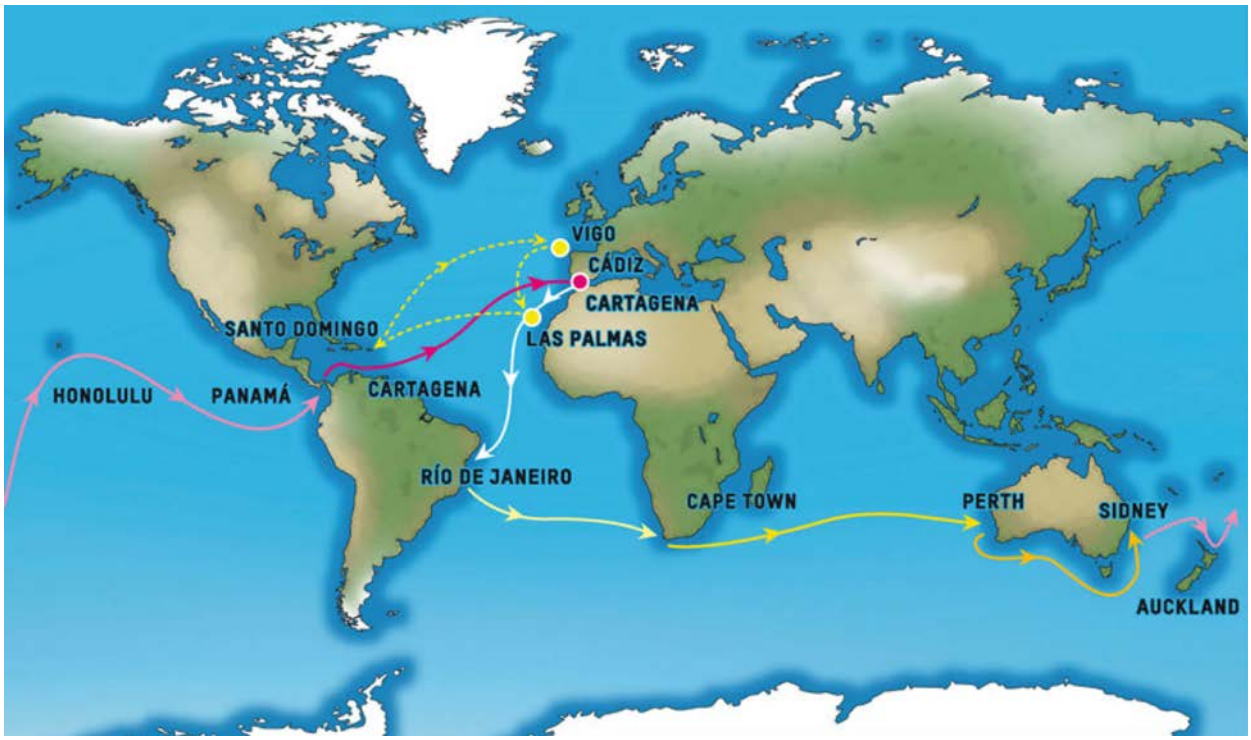


Figura 8. Itinerario de la expedición Malaspina 2010, incluyendo las rutas de la nave Hespérides (línea continua) y del Sarmiento de Gamboa (línea punteada) .

A nivel batipelágico hay una disminución general de la concentración de COD a lo largo de la circulación termohalina profunda (Figura 6). La concentración de COD es alta (> 50 µmol kg^{-1} C) en aguas del Atlántico norte recién formadas (al norte de 50 ºN), tiende a ser un poco más baja y constante en las regiones ecuatoriales (aproximadamente 45 µmol kg^{-1} C) y desciende nuevamente en el sur a un mínimo de 39 µmol kg^{-1} C. La concentración constante de COD a lo largo del sur del Océano Índico (40 µmol kg^{-1} C) sugiere una entrada neta de carbono, debida a la invasión de agua profunda circumpolar (CDW), y una eliminación posterior. La concentración de carbono orgánico en aguas profundas del océano Pacífico va disminuyendo gradualmente a medida que estas avanzan hacia el norte: el COD es de 42 µmol kg^{-1} C en las aguas circumpolares del Pacífico sur

y disminuye a 36 µmol kg $^{-1}$ C (Figura 7, Hansell *et al*. 2009). Probablemente esta disminución del COD se explica por el consumo biológico debido a procariotas heterotróficos y tal vez a los hongos .

Las características abióticas del océano profundo definen un hábitat muy diferente de la superficie. La información sobre la abundancia y la distribución de los microeucariotas en la columna de agua del océano oscuro es fragmentaria (Tanaka y Rassoulzadegan 2002, Yamaguchi *et al*. 2004, Fukuda *et al*. 2007, Sohrin *et al*. 2010, Morgan -Smith *et al*. 2011, Morgan - Smith *et al* . 2013). Se han realizado estudios de diversidad en la columna de agua (López-García *et al* . 2001, Stoeck *et al*. 2003, Countway *et al*. 2007, Not *et al*. 2007), en los sedimentos (Edgcomb *et al* . 2011a, Salani *et al* . 2012) y en las chimeneas marinas (Edgcomb *et al*. 2002, Sauvadet *et al*. 2010). A partir de estos estudios sabemos que la diversidad de los protistas en aguas profundas aparece dominada por Alveolata y Radiolaria mientras que los hongos predominan en los sedimentos. A pesar de no ser uno de los grupos dominantes, Excavata aparentemente prefiere las aguas profundas más que las superficiales. Siendo la fototrofía imposible en la oscuridad del océano profundo, la comunidad batipelágica de los protistas sobrevive gracias a uno de los tres estilos de vida heterótrofos mencionados anteriormente: fagotrofia, osmotrofia y parasitismo. La importancia relativa de cada modalidad trófica en el ecosistema es todavía tema de debate en la comunidad científica.

Para lograr una visión global del funcionamiento de las profundidades del océano, se realizó la campaña oceanográfica *Malaspina* en el año 2010, a bordo del BIO_Hespérides. La campaña se inició en Cádiz (España) y se tomaron muestras de 147 estaciones distribuidas por todo el mundo (Figura 8). El objetivo principal de esta expedición era el estudio del océano profundo a escala mundial, incluyendo la recopilación de datos sobre los microeucariotas. La magnitud y la multidisciplinariedad del esfuerzo de muestreo nos ha permitido comparar los datos con otros paráme-

tros, abiótico y biótico, con el fin de lograr una visión más completa de todo el sistema.

Objetivo de la tesis

El objetivo general de esta tesis es conseguir una visión global de la comunidad de microeucariotas marinos. El logro de este objetivo se estructura en cuatro capítulos. El primer capítulo (*Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean*, ISME 2011), trata sobre el estudio de la diversidad de la comunidad epipelágica mediante bibliotecas de clones, ha sido útil como una primera aproximación a las herramientas de la biología molecular y para establecer unas directrices sobre cómo tratar los conjuntos de datos de secuencias (la alineación, la agrupación, las estimaciones de diversidad). En el segundo capítulo (*General patterns of diversity in major marine microeukaryote lineages*, PLOSONE 2013) las secuencias derivadas de todos los trabajos publicados antes del año 2010 fueron analizadas con el fin de describir las diferentes características de la diversidad genética de los grupos microeucarióticos. Además, se realizó un estudio exploratorio del modelo evolutivo de los diferentes taxones. El fruto más precioso de este trabajo fue un conjunto de secuencias fiables, todas pertenecientes a la región V4 de 18S rADN, que fueron el núcleo de una base de datos de referencia (MAS9013) utilizada para la identificación taxonómica y la búsqueda de quimeras en los estudios sucesivos realizados por pirosecuenciación. La segunda parte de la tesis, en el marco del proyecto Malaspina-2010, se ha centrado en el ecosistema del océano profundo. El tercer capítulo (*Global abundance of planktonic heterotrophic protists in the deep ocean*, enviado a ISME J.) investiga la abundancia de flagelados heterotróficos, en las regiones globales meso y batipelágicas, con el uso combinado de la microscopía de epifluorescencia y la citometría de flujo. En el cuarto capítulo (*Diversity of marine microeukaryotes in the global deep ocean*, in preparación), se estudió la diversidad filogenética y biogeográfica de los microeucariotas, y su relación con los parámetros ambientales en la frontera entre las regiones batipelágica y abisales, a través de la pirosecuenciación de rADN y de

la metagenómica .

Los diferentes temas estudiados se pueden explicar en virtud de dos objetivos generales y varios específicos:

Objetivo 1 : *Definición de los grupos taxonómicos de los microeucariotas marinos y su estructura genética*

La primera parte de la tesis representa un esfuerzo por definir nuestra "unidad de la diversidad" a partir de los estudios basados en la clonación molecular y la secuenciación de Sanger, con el fin de establecer una base sólida para la segunda parte de la tesis. Empezamos con los datos de una campaña oceanográfica (Capítulo 1) y luego continuamos con el análisis de la base de datos de 18S rADN completa disponible en ese momento (Capítulo 2). Los objetivos específicos de esta parte fueron:

- Seleccionar la región del gen 18S rADN que mejor representa la variabilidad del gen completo

- Identificar un umbral de similitud razonable para el agrupado por OTU

- Establecer la distancia máxima en grupos taxonómicos a nivel de clase

- Destacar las clases taxonómicas típicas que forman las comunidades de superficie

Objetivo 2: *Estudio descriptivo de las comunidades del océano profundo global*

La expedición Malaspina nos permitió tener un amplio conjunto de muestras procedentes de todas partes del mundo con los parámetros abióticos y bióticos asociados. Una cantidad tan grande de

datos fue la base para el estudio de los microeucariotas de profundidad (Capítulos 3 y 4) siguiendo los siguientes objetivos específicos:

- Determinar la abundancia, biomasa y distribución de los microeucariotas en la columna de agua entre 200 y 4000 m de profundidad

- Estudiar la diversidad de microeucariotas batipelágicos mediante pirosecuenciación y metagenómica

- Identificar los parámetros abióticos y bióticos que explican la abundancia y la diversidad de los microeucariotas de profundidad, con un énfasis particular en la relación con los procariotas

Capítulo 1

A pesar de la importancia de los protistas marinos que se encuentran en la base de la cadena trófica marina, su diversidad filogenética ha sido poco estudiada, especialmente la de las células más pequeñas que son difícilmente distinguibles por sus características morfológicas. Los avances recientes obtenidos mediante la aplicación de técnicas moleculares en el estudio de la ecología de los protistas han revelado que los ensamblajes de protistas marinos están representados por grupos filogenéticos distantes e incluyen muchos taxones que son nuevos para la ciencia y han sido ignorados y han pasado desapercibidos hasta hace muy poco. En este trabajo se utilizan bibliotecas de clones del Océano Índico publicadas recientemente, con un total de 500 secuencias de 18S ADNr con alrededor de 800 pb, para descubrir el número de diferentes linajes filogenéticos y el número de OTUs (Unidades taxonómicas operativas) observados y estimados. Además ha sido indagada la novedad del conjunto de secuencias de datos. Este análisis cuantifica la magnitud de la diversidad, a diferentes escalas filogenéticas y la novedad de la señal molecular obtenida desde el plancton microbiano marino. El alto nivel de diversidad y novedad detectada tiene claras implicaciones evolutivas y ecológicas.

Capítulo 2

Los microeucariotas tienen un papel vital para el funcionamiento de los ecosistemas marinos, pero aún así algunas características generales de su diversidad y filogenia siguen sin estar claras. En este trabajo se investigaron dos aspectos de los principales linajes de microeucariotas oceánicos utilizando secuencias de 18S ADNr (las regiones hipervariables V4-V5) provenientes de bases de datos públicas y que derivan de diferentes estudios ambientales marinos. Se generó así un conjunto de datos manualmente curados de 8291 secuencias de Sanger y posteriormente se dividieron en 65 grupos taxonómicos (más o menos al nivel de clase, basándose en KeyDNATools) antes de los análisis. En primer lugar , se calcularon las distancias genéticas y el numero de secuencias agrupadas en unidades taxonómicas operativas (OTUs) utilizando diferentes niveles de umbral de distancia. Se encontró que la mayoría de los grupos taxonómicos tenían una distancia genética máxima de 0.25. En segundo lugar, se utilizaron los árboles filogenéticos para estudiar los patrones evolutivos generales. Estos árboles confirmaron nuestra clasificación taxonómica y sirvieron para determinar la evolución de los linajes a través del tiempo (LTT). Los resultados de LLT indicaron diferentes dinámicas de cladogénesis entre los grupos, con algunos mostrando una diversificación temprana en su historia evolutiva y otros una más reciente. En general, nuestro estudio proporciona una descripción mejorada de la diversidad de microeucariotas en los océanos en términos de diferenciación genética dentro de los grupos, así como en la estructura filogenética general. Estos resultados serán importantes para interpretar la gran cantidad de datos de las secuencias que se van generando actualmente por las tecnologías de secuenciación de alto rendimiento.

Capítulo 3

El océano oscuro es uno de los más grandes biomas de la Tierra, con un papel crítico en la remineralización de la materia orgánica y en el secuestro de carbono a nivel mundial. A pesar de su reconocida importancia, poco se sabe acerca de la comunidad de los protistas heterótrofos (HP), que son probablemente los principales consumidores de biomasa procariota. Para investigar este componente microbiano a escala global, a fin de determinar su abundancia y biomasa en aguas meso y batipelágicas en las muestras de la circunnavegación Malaspina-2010, se ha usado una combinación de la microscopía de epifluorescencia y la citometría de flujo. Los HP eran claramente omnipresentes en el océano profundo global, incluso en las muestras más profundas investigadas (4000 m). Su abundancia disminuía con la profundidad, de un promedio de 72 ± 19 células mL^{-1} en aguas mesopelágicas a 11 ± 1 células mL^{-1} en aguas batipelágicas, mientras que su biomasa mundial disminuyó de 280 ± 46 pg C mL^{-1} a 50 ± 14 pg C mL^{-1}. Los parámetros que mejor explican la varianza de la abundancia de HP son la profundidad y la abundancia de procariotas, y en menor medida el oxígeno y los virus de genoma grande. Diferentes señales sugirieron la depredación activa por parte de HP en procariotas, tales como la presencia de flagelos en la mayoría de las células y la generalmente buena correlación con la abundancia procariota. El ratio entre la abundancia de procariotas y la de HP variaba a escala regional y los sitios con valores mayores aparecen relacionados con una mayor contribución de organismos osmotrofos en la comunidad. Nuestro estudio permite una mejor comprensión de la relación entre HP y su entorno, arrojando luz sobre su importancia como actores en la red trófica microbiana del océano oscuro .

Capítulo 4

El objetivo de este trabajo es estudiar la diversidad de los microeucariotas batipelágicos. Las muestras de agua de mar (desde 3000 hasta 4000 m de profundidad) procedían de 27 estaciones de la expedición global Malaspina-2010 incluyendo el Atlántico, Pacífico e Índico. Se utilizó la pirosecuenciación para obtener más de medio millón de secuencias de la región V4 del 18S ADNr y después de varias etapas de curación (eliminación de las secuencias de baja calidad, más cortas de 250 bp, presentes en una sola muestra y quimeras) se agruparon en 2482 OTU al 97% de similitud. La abundancia relativa de las 20 OTU más abundantes coincide adecuadamente con los resultados de un análisis paralelo de metagenómica, lo que sugiere que la producción de secuencias se vio poco afectada por los sesgos de la técnica de la PCR. Apareció una tendencia débil de similitud genética entre estaciones geográficamente cercanas y entre muestras de la misma masa de agua. Además, el ratio de abundancia de células entre procariotas y microeucariotas tenía una relación significativa con la composición taxonómica. A pesar de que 42 OTUs se encontraron en todas las muestras, no se encontró una comunidad global típica. En cambio, aparecieron cuatro grupos filogenéticos principales (Collodaria, Crisófita, MALV-II y Basidiomycota) presentes en diferentes proporciones en cada lugar. La cantidad de novedad filogenética se concentra en tres puntos, uno en cada océano y representa el 6 % de la secuencias globales. Las curvas de rarefacción señalan que aún hay especies por descubrir. Nuestro estudio es el primer paso esencial hacia una investigación más detallada de la microbiota del océano profundo.

Síntesis de los resultados y discusión general

El primer objetivo de esta tesis es mejorar la información sobre la estructura genética de los grupos de microeucariotas, principalmente a nivel de clase, con el objetivo final de construir una base de datos de secuencias depuradas y fiables. Una vez definida esta referencia, procedente sólo de los estudios moleculares realizados mediante secuenciación de Sanger, fue utilizada como marco para estudiar la diversidad global de los microeucariotas en las profundidades del océano usando secuenciación de alto rendimiento. La diversidad de microeucariotas de profundidad , más los datos recogidos de su abundancia y biomasa , permitió hacer un dibujo refinado del medio ambiente batipelágico .

El reto de la taxonomía molecular: herramientas para definir un grupo

La región V4 del 18S rDNA : el mejor compromiso

La gestión del gran número de secuencias provenientes de los estudios moleculares necesita un gran esfuerzo inicial para establecer buenos criterios de trabajo. Las secuencias deben ser organizadas en categorías (OTU) definidas por determinados niveles de similitud con el fin de obtener valores cuantitativos de diversidad. El primer paso de nuestro trabajo fue definir la región diana del 18S ADNr, ya que actualmente la tecnología de secuenciación no está lista para analizar el gen completo. En particular, la elección era entre las regiones V4 y V9. Este debate nació dentro de "BioMarKs", un proyecto europeo de investigación que estudia la diversidad de protistas con técnicas moleculares y microscópicas . Estudios seminales que utilizaban pirosecuenciación 454 se centraron en la región V9 (Amaral-Zettler *et al.* 2009, Cheung *et al*. 2009, Stoeck *et al*. 2009) que es una región muy corta (cerca de 150 pb) y fue óptima para la tecnología de secuenciación

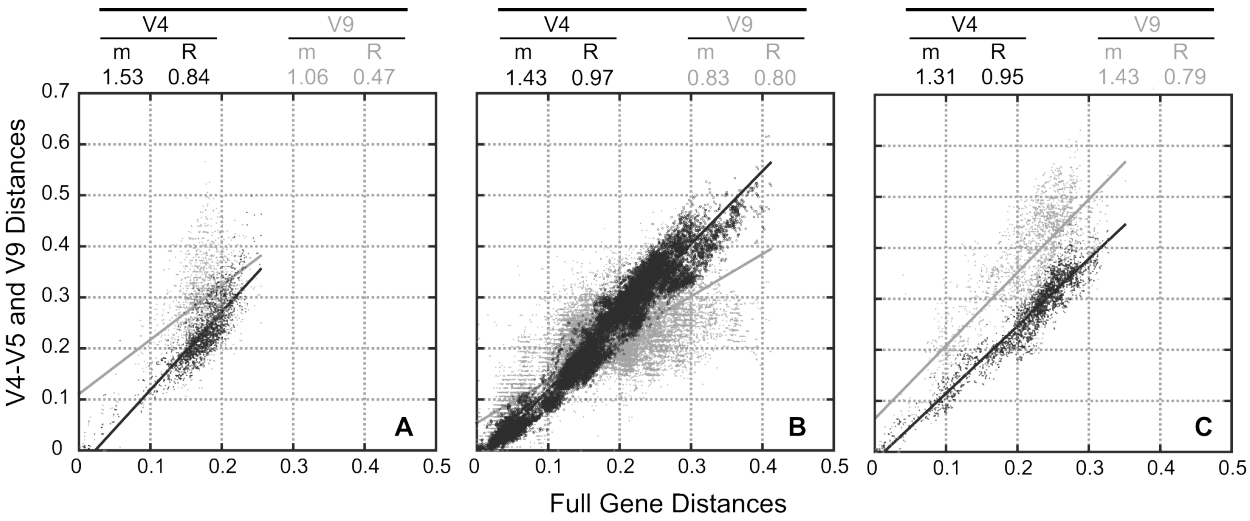


Figura 1, Capítulo 2. Comparación de secuencias parciales y completas de 18S ADNr para inferir distancias genéticas. Los tres paneles muestran las distancias genéticas (Jukes Cantor corregida) del gen completo en relación a zonas parciales (V4-V5 en gris oscuro o V9 en gris claro) para secuencias que pertenecen a Stramenopiles (A), Alveolata (B), y Rhizaria (C). Las pendientes (m) y los coeficientes (R) de las correlaciones se muestran en la parte superior de los gráficos.

en ese momento. Gracias a los avances técnicos se produjeron secuencias más largas (actualmente hasta 400 pb) y fue entonces posible analizar regiones hipervariables, como la V4. Hemos contribuido a este debate mediante la búsqueda de la región que mejor representa el gen entero. Nuestros resultados sugieren que la variabilidad detectada en la región V4 (alrededor de 500 pb) es un buen indicador de la variabilidad que se observa en el análisis de todo el gen. Las pendientes de las líneas de regresión (Figura 1 ,Capítulo 2) entre las distancias calculadas con el gen y la región V4 eran alrededor de 1.4 para tres supergrupos diferentes, por lo que las distancias calculadas con la región V4 se podrían traducir a distancias para todo el gen dividiendo por este valor.

Cómo definir una clase taxonómica : desde la filogenia a la agrupación

La tecnología de secuenciación mejora rápidamente, tal vez más rápido que nuestra capacidad para gestionarla. Los métodos de clonación molecular generalmente producen entre 100 y 500

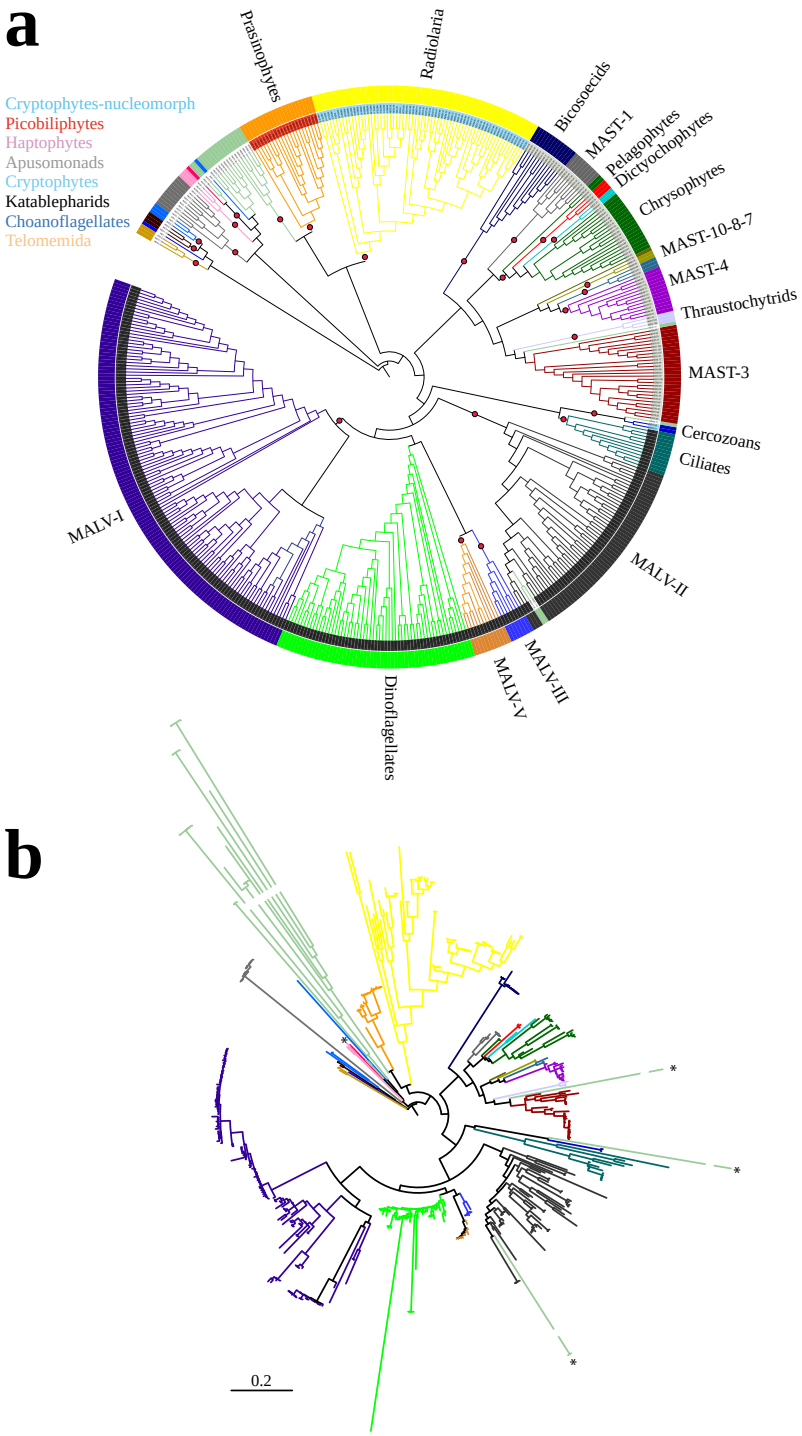


Figura 1, Capítulo 1. Árbol filogenético de secuencias de 18S ADNr de picoeucariotas muestreadas en el océano Índico. El árbol fue construido con 500 secuencias de ~ 815 bases en 961 posiciones. (a) Los colores del árbol representan el supergrupo eucariota (anillo interior) y el grupo taxonómico determinado (ramas de colores y el anillo exterior; se muestran los nombres). Las ramas que conducen a los grupos con valores de "bootstrap" por encima del 70% están marcados con un punto rojo. (b) El mismo árbol que muestra las longitudes de rama, con los mismos colores que el anterior. La longitud de las nuevas OTU (verde claro) se ha reducido a la mitad y las que se han colocado filogenéticamente están marcadas con un asterisco. La barra de escala indica 0,2 sustituciones por posición.

secuencias por muestra. A pesar del hecho de que este número es muy bajo en comparación con el rendimiento de la pirosecuenciación , el proceso global está más controlado y las secuencias obtenidas son más largas. Por lo tanto, las secuencias producidas con bibliotecas de clones son muy útiles para construir una alineación fiable, que es la base para una filogenia correcta, como hemos hecho en el primer capítulo (Figura 1). Durante la «era de las bibliotecas de clones» la construcción de un árbol era la mejor manera de definir un grupo taxonómico, pero hoy en día la superproducción de secuencias hace esta operación más difícil, especialmente el alineado de un gran número de secuencias genéticamente distantes. Ciertamente, las secuencias cortas no pueden resolver las relaciones taxonómicas entre linajes distantes. En la «era de la pirosecuenciación « hay un cambio progresivo desde los árboles filogenéticos hacia la *agrupación* mediante la cual las secuencias se agrupan en unidades taxonómicas sobre la base de similitud de secuencias. La agrupación de las secuencias en un árbol se basa en distancias patrísticas entre las secuencias (longitudes de rama), que dependen de la cantidad y variabilidad de las secuencias consideradas y comparan cada secuencia con todas las demás. Por otra parte , la *agrupación* en OTU se basa en la similitud (o distancias genéticas corregidas tales como Jukes Cantor) entre pares de secuencias. El valor absoluto de similitud o distancia es independiente del número de secuencias consideradas ya que se realizan comparaciones sólo por pares. La comunidad científica está adoptando como rutina la agrupación de secuencias basadas en la similitud y el problema es que en este proceso se pierde la forma del árbol, por lo que ahora es más difícil de identificar valores atípicos y linajes que evolucionan rápidamente. En el primer capítulo comprobamos si el número de OTU de los dos enfoques era diferente y encontramos que hasta distancias de 0.10, la agrupación utilizando distancias JC (casi equivalentes a la similitud) o distancias patrísticas (desde el árbol filogenético) mostró buena correspondencia (Figura 2, Capítulo 1). Teniendo en cuenta que a menudo el

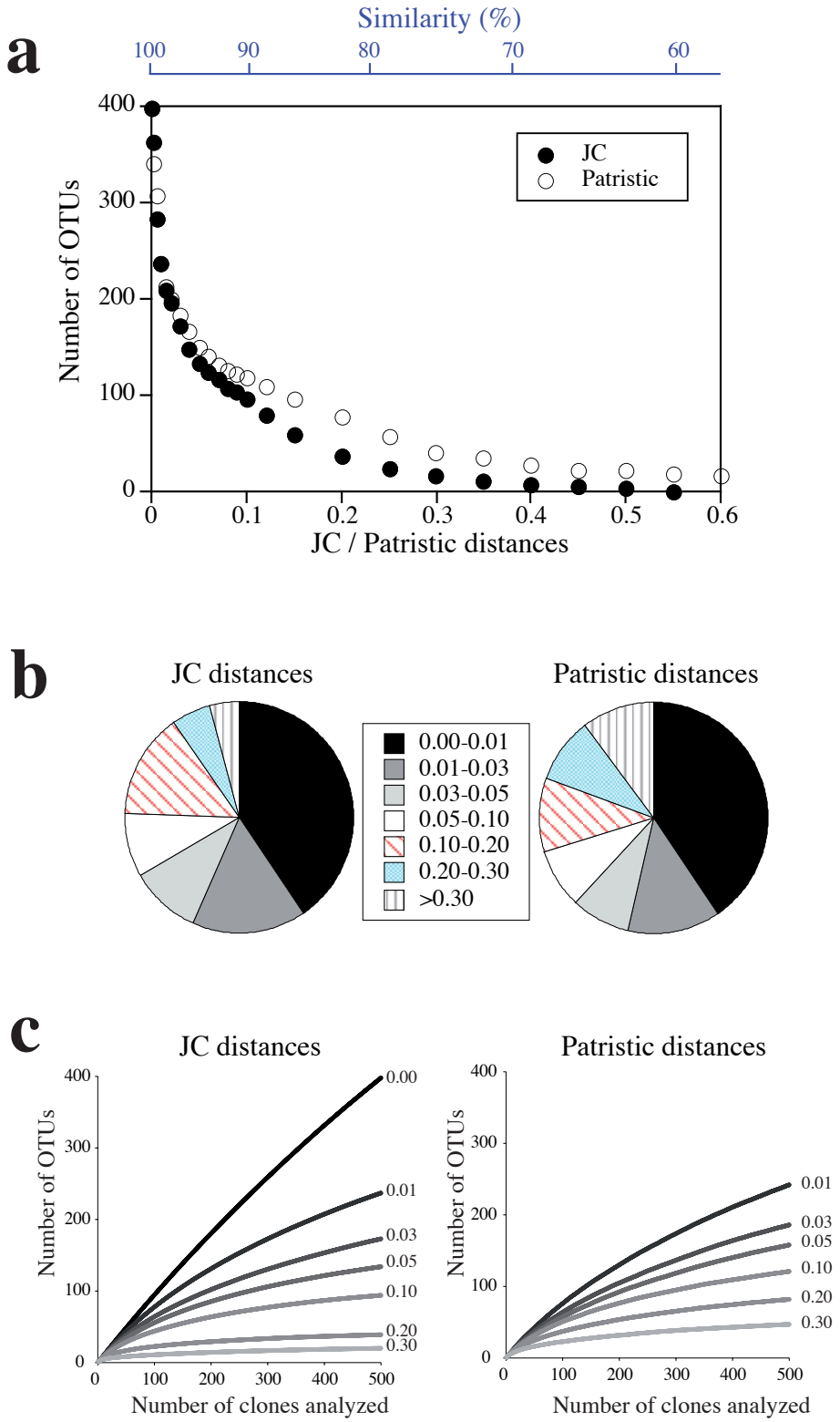


Figura 2, Capítulo 1. (a) Número de OTU observado después de agrupar las 398 secuencias únicas del Océano Índico en los diferentes niveles de agrupamiento basado en distancias Jukes-Cantor o patrística. La correspondencia entre la distancia Jc y la similitud entre secuencias se muestra en la parte superior de la gráfica para propósitos comparativos. (b) Distribución del número de OTU en clases de distancia para ambos métodos de agrupación. El área en cada clase representa la diferencia en las OTU observadas en los dos límites de la clase (así las OTU disminuyen al relajarse las condiciones de agrupamiento entre los dos límites). (c) Las curvas de rarefacción (OTU observadas versus clones analizados) a niveles discretos de distancia de agrupación (desde 0.00 a 0.30) para ambos métodos de agrupación.

agrupamiento se realiza al 95 -97 % de similitud, este se puede considerar un resultado aceptable. A pesar de que las distancias patrísticas son la base para una agrupación evolutiva más sólida y precisa, la agrupación de similitud se sigue utilizando con conjuntos de datos de 454, ya que se omite la etapa de la alineación y además proporciona una manera simple e intuitiva para analizar un número elevado de secuencias .

Definiendo los límites de los grupos taxonómicos

Cuando se trabaja con el método de la agrupación es esencial poder responder a dos preguntas: ¿cuál es el umbral de distancia que define una OTU con significado biológico útil (es decir, una especie)? y ¿cuál es la distancia máxima que se puede encontrar dentro de un grupo determinado?. Acerca de la primera pregunta, diferentes autores han propuesto diferentes niveles de umbral(Worden 2006, Jeon *et al*. 2006 , Caron *et al*. 2009), pero hay poco apoyo para justificar cada hipotético nivel y sería importante que el mundo científico llegara a una conclusión en esta dirección. Debido a los polimorfismos intragenómicos (Introducción, Tabla 1) y a los errores de secuenciación pensamos que no es aconsejable el uso del 100% de similitud para definir las OTU. Un valor entre 97-99% de similitud parece ser un criterio más razonable, ya que es suficiente para ser bastante estricto, pero no tanto como para separar secuencias que difieren a causa del polimorfismo intragenómico o por errores de secuenciación. En el capítulo dos se abordó la segunda pregunta para los grupos más o menos equivalentes a las clases taxonómicas en la sistemática clásica, incrementada por la falta de la representación gráfica del árbol que hizo necesaria una forma alternativa para identificar rápidamente las secuencias de valores atípicos. Se realizaron árboles para comprobar la afiliación de las secuencias y luego fueron analizadas por grupo. Encontramos que el 75% de los grupos a nivel de clase tenían una distancia genética máxima (corregida) por debajo de 0.25. Esta es ahora nuestra referencia general para la distancia máxima permitida dentro de una

Tabla 1, Capítulo 2. Clasificación de las secuencias de 18S ADNr ambientales en 42 grandes grupos taxonómicos. Cada grupo se codifica de acuerdo a su rango taxonómico (S: suborden; C: clase; O: orden; G: género; R: ribogrupo). La tabla muestra el número de secuencias por grupo (SEC), la media (AVG), máximo (Max) y máximo corregido (MAXC) de distancias y el número de OTU en tres niveles de corte.
* Nassellaria comprende también el orden Collodaria

Supergroup	Group			Distances			OTUs		
			Seq	Avg	Max	Max_c	0.00	0.01	0.05
Opisthokonta	*Choanoflagellatea*	C	100	0.13	0.30	**0.24**	89	**56**	32
Rhizaria	*Acantharea*	C	129	0.15	0.29	**0.26**	110	**63**	29
	Chlorarachniophyceae	C	33	0.14	0.24	**0.23**	29	**13**	7
	Larcopyle	O	18	0.02	0.05	-	13	**4**	1
	Monadofilosa	S	81	0.11	0.30	**0.22**	72	**56**	33
	*Nassellaria**	O	52	0.18	0.41	**0.32**	45	**29**	19
	RAD A	R	37	0.17	0.29	**0.26**	34	**23**	15
	RAD B	R	88	0.11	0.23	**0.16**	66	**36**	17
	Spumellaria	O	209	0.06	0.26	**0.13**	154	**79**	20
Archaeplastida	*Prasinophyceae*	C	551	0.09	0.31	**0.21**	376	**130**	30
	Trebouxiophyceae	C	89	0.01	0.12	**0.04**	26	**11**	6
Stramenopiles	*Bacillariophyceae*	C	253	0.14	0.30	**0.29**	207	**120**	57
	Bicosoecea	C	75	0.11	0.35	**0.28**	60	**34**	17
	Bolidophyceae	C	63	0.05	0.12	**0.11**	34	**12**	7
	Chrysophyceae	C	152	0.13	0.27	**0.24**	115	**75**	32
	Dictyochophyceae	C	91	0.09	0.22	**0.16**	65	**35**	16
	Eustigmatophyceae	C	15	0.01	0.03	-	11	**3**	1
	Labyrinthulida	C	29	0.17	0.35	**0.34**	26	**19**	17
	MAST-1	R	107	0.08	0.20	**0.16**	74	**28**	9
	MAST-2	R	20	0.01	0.05	-	13	**6**	2
	MAST-3	R	149	0.12	0.27	**0.21**	110	**73**	31
	MAST-4	R	92	0.03	0.07	**0.06**	60	**24**	3
	MAST-7	R	82	0.04	0.14	**0.08**	48	**21**	6
	MAST-8	R	17	0.07	0.13	-	14	**9**	6
	MAST-12	R	26	0.16	0.27	-	24	**19**	16
	Oomyceta	C	19	0.11	0.29	-	16	**13**	10
	Pelagophyceae	C	34	0.01	0.07	**0.02**	22	**8**	2
	Pirsonids	-	47	0.03	0.09	**0.08**	37	**26**	5
CCTH	*Cryptophyceae*	C	179	0.09	0.24	**0.21**	130	**45**	3
	Katablepharids	-	20	0.02	0.06	-	12	**6**	2
	Picobiliphyceae	R	53	0.07	0.20	**0.15**	42	**24**	8
	Prymnesiophyceae	C	193	0.08	0.30	**0.14**	148	**90**	37
	Telonemia	C	68	0.05	0.12	**0.11**	60	**42**	9
Alveolata	*Ciliophora*	P	956	0.18	0.42	**0.37**	788	**434**	187
	Dinophyceae	C	1018	0.07	0.50	**0.24**	848	**463**	122
	MALV-I	R	980	0.19	0.48	**0.42**	779	**431**	132
	MALV-II	R	1815	0.16	0.38	**0.30**	1517	**900**	353
	MALV-III	R	79	0.05	0.15	**0.11**	60	**38**	9
	MALV-V	R	51	0.02	0.07	**0.04**	41	**19**	3
Excavata	*Diplonemea*	C	58	0.11	0.21	**0.21**	56	**51**	27
	Kinetoplastea	C	40	0.23	0.39	**0.37**	31	**22**	15
Incertae sedis	*Apusomonadidae*	C	14	0.15	0.41	-	9	**6**	4

clase y es un valor útil para interpretar la equivalencia taxonómica de ribogroupos ambientales.

La importancia de una base de datos de referencia

Una buena base de datos de referencia es una herramienta esencial para cualquier estudio molecular basado en secuencias cortas, pero las bases de datos actuales comprenden sólo los procariotas (Greengenes) o dan un tratamiento menos preciso a los eucariotas (SILVA). El principal problema de SILVA es que a menudo carece de una buena asignación taxonómica de las secuencias de eucariotas, especialmente de los ribogrupos recién descubiertos. Una nueva herramienta para la identificación de eucariotas a partir del 18S ADNr se ha publicado muy recientemente, la base de datos PR2 (Base de datos Protista ribosomal de referencia, Guillou *et al.* 2013). Esta herramienta no existía al comienzo de esta tesis. En este marco, quiero destacar la importancia del conjunto de secuencias depuradas del capítulo dos (8291 secuencias), que constituye el núcleo, mejorado con PR2, de una base de datos interna de referencia (MAS9013), que se ha utilizado para la asignación taxonómica y detección de quimeras en la segunda parte de la tesis y en otras publicaciones en preparación.

Composición típica de la comunidad de microeucariotas epipelágicos

La diversidad observada en el capítulo 1, en términos de abundancia relativa de determinados linajes (Figura 1), es la típica que se encuentra en otros estudios moleculares de picoeucariotas marinos (Massana y Pedrós-Alió 2008, Vaulot *et al.* 2008) y se asemeja a la abundancia de grupos en el capítulo 2 (Tabla 1). Los alveolados, principalmente MALV-I y MALV-II, dominan la comunidad y representan el 47% de los clones, seguidos por los estramenopilos (19%) y Rhizaria (13%). Los hongos no fueron considerados en los dos primeros capítulos, ya que generalmente están poco representados en el entorno epipelágico, a nivel mundial comprenden menos del 1%

de las secuencias de las bibliotecas de clones (Massana y Pedrós-Alió 2008). Las diferencias en la composición taxonómica entre epipelágico y ecosistema profundo, incluso a niveles de supergrupos, son evidentes (Figura 6, Capítulo 4) y serán analizadas en la segunda parte de esta discusión. Curiosamente, a bajas distancias (mínimo 1300 km), las muestras diferían fuertemente cuando se analizaban por bibliotecas de clones (Figura 6, Capítulo 1) y en ese momento esto fue explicado por un efecto del submuestreo. Sin embargo, a distancias comparables y con un gran esfuerzo de secuenciación, vemos que las diferencias entre las muestras profundas también están presentes (Figura 5, Capítulo 4). Esta fue la primera señal del fuerte efecto del medio ambiente en la selección de la comunidad.

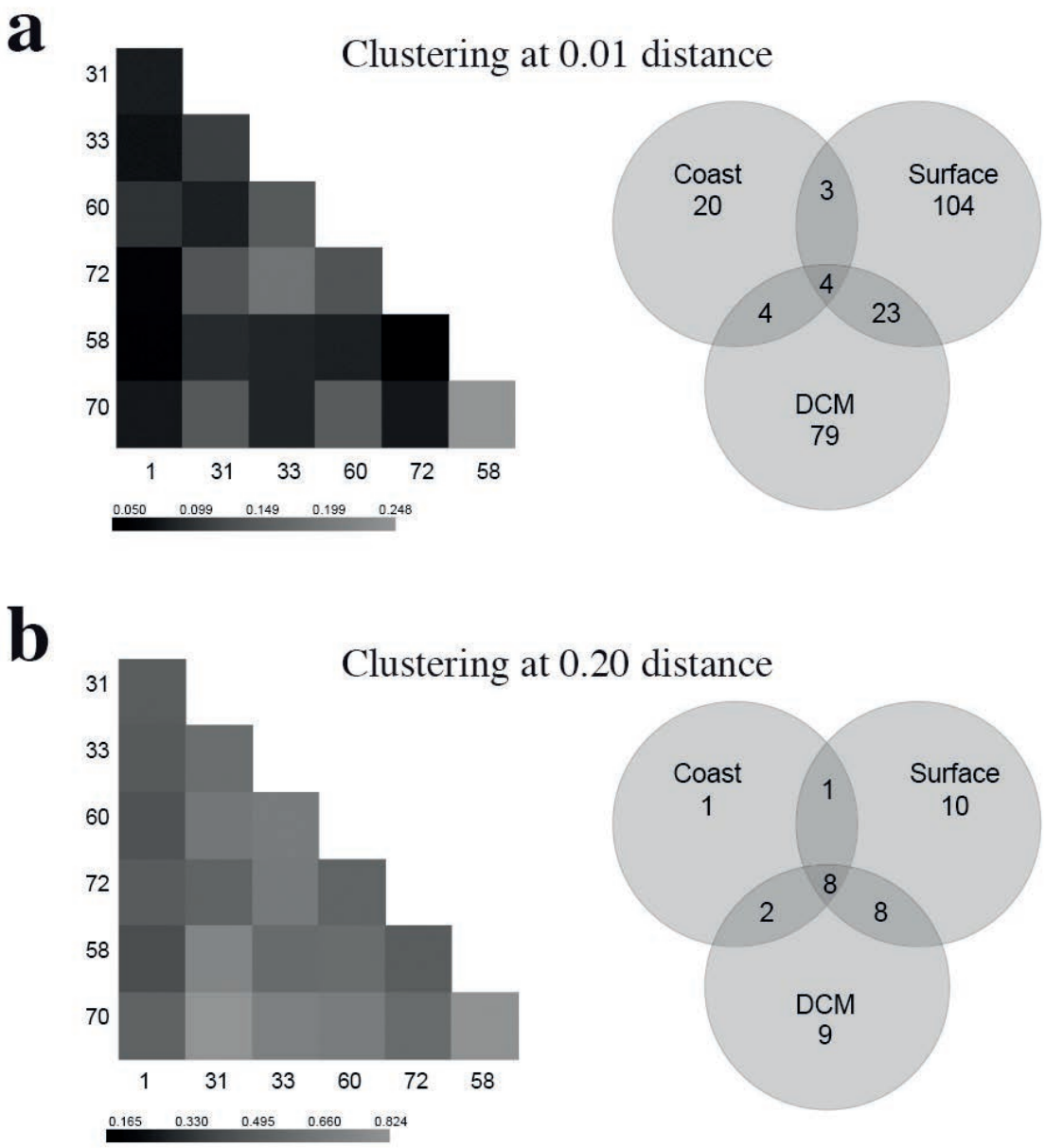


Figura 6, Capítulo 1. "Heatmaps" (izquierda) y diagramas de Venn (derecha) que comparan la diversidad de picoeucariotas marinos entre las muestras, utilizando las OTU compartidas o exclusivas definidas en la agrupación de distancias JC de 0.01 (a) o 0.20 (b). Las muestras son costeras (1), de alta mar superficiales (31, 58 y 70) o de DCM (33, 60 y 72).

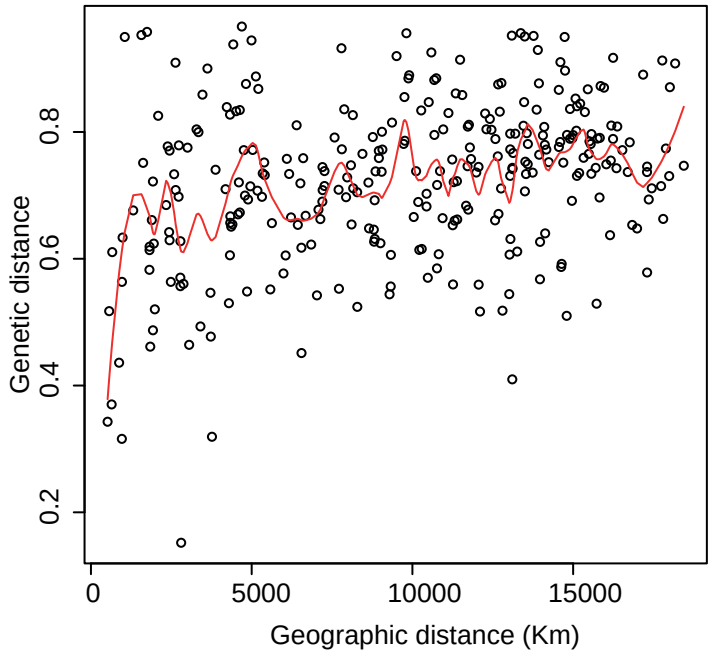


Figura 5, Capítulo 4. Test de Mantel de la relación entre las distancias genéticas (Bray-Curtis) de los conjuntos de protistas profundos y la distancia geográfica entre las muestras. También se muestra la línea de la interpolación de los valores (línea LOESS).

El océano profundo

Contando microeucariotas : la necesidad de la citometría de flujo

Por lo que sabemos este es el primer estudio que aplica la citometría de flujo, junto con la microscopía, en una investigación a gran escala (Capítulo 3). La microscopía de epifluorescencia es un trabajo que necesita mucho tiempo y es propensa a errores del operador, mientras que la citometría de flujo presenta otros tipos de problemas. Es posible identificar varias poblaciones de microeucariotas fotosintéticos gracias a sus pigmentos (Olson *et al*. 1993, Li *et al.* 1994, Marie *et al.* 2000), pero para detectar microeucariotas heterotróficos se requiere una coloración (en este caso SYBR Green). Sin embargo, esto no discrimina entre procariotas y eucariotas y puede existir una continuidad entre eucariotas de pequeño tamaño y grandes bacterias. Para resolver este problema, se utilizó la microscopía de epifluorescencia con el fin de posicionar la ventana de conteo en el software de citometría en muestras seleccionadas (a continuación, se aplicó al conjunto de datos completo) y para comprobar los valores de estaciones con abundancias poco realistas. Un método alternativo para ahorrar tiempo y recuperar el tamaño y la forma de las células, sería la microscopía de epifluorescencia automática, que se aplicará en el futuro en nuestro laboratorio. La comparación de la citometría de flujo y los recuentos microscópicos (Figura 2, Capítulo 3) fue muy buena ($R^2 = 0.82$, $p < 0.0001$). Por lo tanto, un gran número de muestras fueron procesadas por citometría de flujo. La abundancia de microeucariotas así determinados fue uno de los dos parámetros en la descripción de la comunidad global del océano profundo .

Características generales de los microeucariotas en el océano batipelágico

En la región batipelágica (1000- 4000 m), que fue también objeto del estudio la diversidad, la abundancia de microeucariotas promedio fue de 14 células mL^{-1}. Esta concentración no es cons-

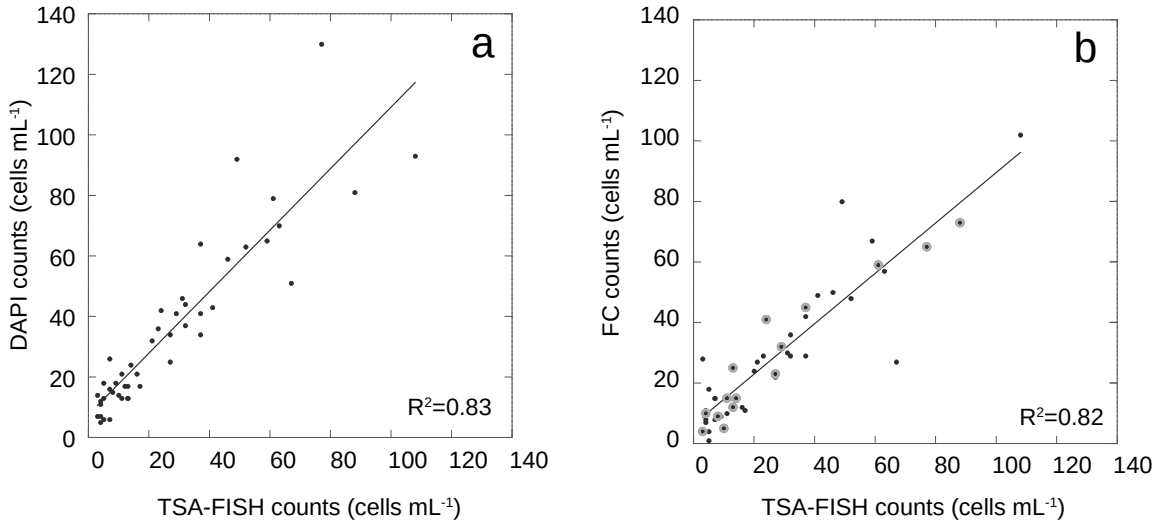


Figura 2, Capítulo 3. Comparación metodológica de recuentos de protistas heterótrofos del océano profundo. Recuentos de DAPI versus recuentos de Citometría de Flujo (FC) (a) y FC en comparación con los recuentos de TSA- FISH (b) en las muestras de diez perfiles verticales seleccionados (que se muestran como estaciones numeradas en la figura 1). Las muestras que se utilizaron para colocar la ventana FC están rodeadas por una zona gris claro en el panel B.

tante, lo cual es particularmente es evidente en el Pacífico Sur, donde hay un pico de 58 células mL^{-1} en la muestra de mayor profundidad de la estación 98. En cuanto a la estructura del tamaño celular, el porcentaje de células muy pequeñas (diámetro equivalente < 3 m) disminuye con la profundidad (Figura 6, Capítulo 3) y a partir de las imágenes tomadas por las mediciones de biomasa, sabemos que algunas de estas células son claramente flageladas. La biomasa promedio de microeucariotas es 50 Pg C ml^{-1} en la capa de 1400-4000 m. Para el estudio de la diversidad de las muestras más profundas (~ 4.000 m) fueron filtrados 120 L, lo que significa que hemos recogido cerca de 1.320.000 células por muestra. La mayoría de las secuencias recuperadas pertenecen a Rhizaria, seguido por Alveolata y Stramenopiles (Figura 6 , Capítulo 4). A nivel local y a un rango taxonómico menor, las comunidades están dominadas fundamentalmente por 3 clases (Collodaria, Crisófitas, y Basidyomicota) y un ribogrupo (MALV -II). Las diferencias en la abundancia y la diversidad entre muestras profundas de microeucariotas están claramente relacionadas, tanto a los

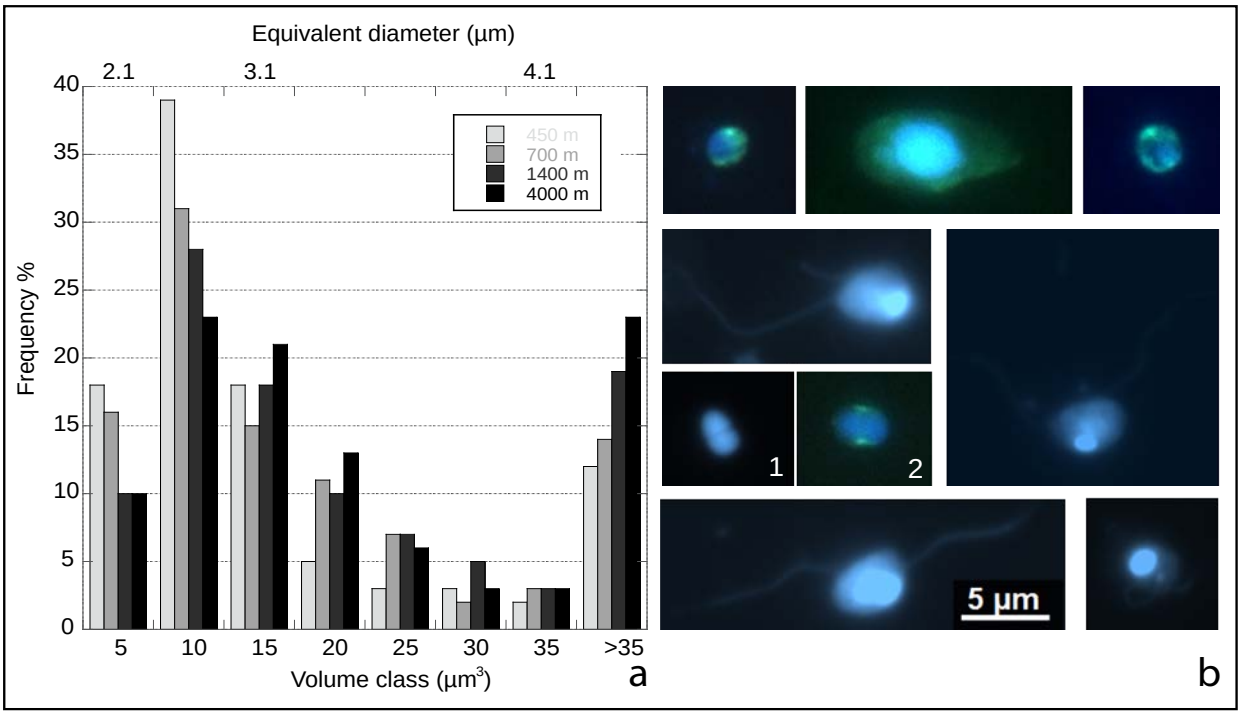


Figura 6, Capítulo 3 (a) Espectros del biovolumen celular de las células HP en diferentes capas de profundidad. También se indica para cada biovolumen celular el diámetro esférico equivalente. (b) Algunas micrografías de células HP batipelágicas, que muestran diferentes formas de células y la presencia de flagelos. La señal azul corresponde al núcleo teñido con DAPI y la señal verde al citoplasma teñido con TSA-FISH. Morfotipos parciales se muestran en las figuras 1 y 2.

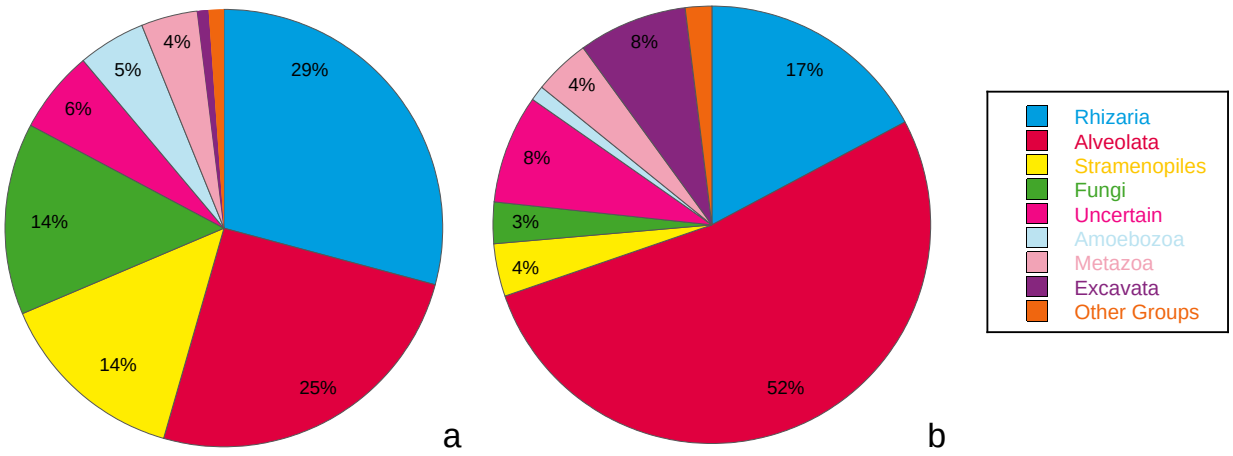


Figura 6, Capítulo 4. Cuadro general de la diversidad de los protistas de profundidad a nivel taxonómico de supergrupo. (a) Número de pyrotags por cada supergrupo, con un promedio de la abundancia relativa de cada muestra. (b) Número de OTU97 por cada supergrupo.

parámetros abióticos (oxígeno, temperatura) como bióticos (abundancia procariótica y viral).

Tabla 3, Capítulo 4. Las veinte OTU más abundantes con su número de secuencias, ocurrencia, identificación taxonómica y la similitud con la referencia medioambiental más cercana (CEM) y la correspondencia mayor con una especie cultivada (CCM).

OTU ID	Pyrotags	OCC	Group	CEM	% SI	CCM	% SI
146	45261	27	*Collodaria*	GU825331	90	*Collophidium ellipsoidae*	90
6539	29099	27	*Basidiomycota*	HQ438183	99	*Tilletiopsis minor*	98
941	23512	27	*Chrysophyceae*	JQ782092	99	*Pedospumella encystans*	98
3736	16125	24	*Spumellaria*	EF172914	99	*Cladococcus viminalis*	96
2627	13645	27	*Dinophyceae*	EU500130	100	*Lepidodinium chlorophorum*	99
309	11748	27	*Ascomycota*	GQ120160	99	*Engyodontium album*	99
1730	11131	27	*Amoebozoa*	GU320596	99	*Platyamoeba contorta*	90
2006	8958	20	*Uncertain*	JX194706	77	*Collozoum Serpentinum*	85
2275	8364	25	*Chrysophyceae*	KC306509	98	*Ochromonas distigma*	97
1165	8228	26	*Collodaria*	GU219126	99	*Collophidium ellipsoidae*	94
2418	8151	27	*Metazoa*	AY937332	99	*Gilia reticulada*	98
7646	6784	27	*Chrysophyceae*	HM749946	99	*Mallomonas Tonsurada*	92
2825	5694	27	*MALV-II*	FN598288	100	*Amoebophyra sp.*	89
6489	5597	25	*Uncertain*	GU824572	82	*Collozoum Serpentinum*	88
4675	4604	25	*Chrysophyceae*	KC306509	98	*Ochromonas distigma*	97
3936	4472	21	*Collodaria*	GU825728	96	*Collophidium ellipsoidae*	96
149	4404	20	*Collodaria*	AY046728	96	*Collophidium ellipsoidae*	84
4324	3565	27	*MALV-II*	JX194526	98	*Amoebophyra sp.*	90
5203	3552	15	*Collodaria*	GU824619	82	*Collozoum Serpentium*	94
7437	2568	27	*Amoebozoa*	FN598227	98	*Platyamoeba contorta*	89

Protistas cultivados en aguas profundas

Sorprendentemente, el análisis de las primeras veinte OTU, que representan juntas el 50% del total de las secuencias, muestran varios ejemplos con una similitud alta con los organismos cultivados (Tabla 3 , Capítulo 4). Esto no es nuevo por lo que se refiere a los hongos (Bass *et al.* 2007, Richards *et al.* 2012), pero fue inesperado para los otros grupos. Por ejemplo, 3 OTU de crisófitas que explican el 79% de las secuencias totales de esta clase son muy similares a las especies cultivadas. Sabemos que en el ambiente epipelágico normalmente existe poco acuerdo entre la diversidad detectada por estudios moleculares y la basada en cultivos (Massana *et al.* 2004), y por lo tanto considerando las características ambientales del océano profundo y su aislamiento relativo, esperábamos encontrar un consenso menor. Sin embargo, un tercio de las OTU más abundantes tienen una similitud mayor al 97% con una especie cultivada y representan el 28% de las secuencias totales. Además dos de estas OTU son en un 99% similares a especies cultivadas. Teniendo

en cuenta que la mayoría de especies cultivadas derivan de muestras epipelágicas, se puede deducir que al menos una cuarta parte de las secuencias están compartidas entre la superficie y las comunidades profundas. Esto demuestra la gran capacidad de adaptación de los microeucariotas a diferentes ambientes. Además, la observación de un impacto diferente del sesgo de cultivo entre comunidades de superficie y de profundidad es un tema que merece ser examinado más a fondo.

Ubicuidad : Todo está en todas partes?

Finlay *et al.* (2004) definió la diferencia entre la *ubicuidad* y la *dispersión ubicua* afirmando que la mayoría de los protistas están caracterizados por la segunda, lo que implica que no necesariamente se encuentran en todas partes, sino que deben estar presentes en los hábitats adecuados en todo el mundo (Caron *et al*. 2009). Otro aspecto importante de este debate es que la presencia de la misma secuencia en océanos separados nos puede dar información acerca de la ubicuidad de las respectivas especies, mientras que su ausencia puede ser simplemente debido al submuestreo. En el conjunto de datos Malaspina, 42 OTU están presentes en todas las estaciones (27) y representan el 80% de las secuencias. Esta visión cualitativa de las comunidades se acerca a la idea de que "todo está en todas partes". Sin embargo, siguiendo el concepto de "el medio ambiente selecciona", la distribución de estas OTU no es uniforme y una OTU que pertenece a la biosfera rara en una muestra a menudo es la OTU dominante en otra, como se ha observado en varios ejemplos, como es el caso de los hongos (Figura 7, capitulo 4). Como se ha visto en la introducción, la homogeneidad ambiental del océano profundo es una vieja y engañosa idea y, a pesar de la supuesta capacidad de dispersión alta, hay un fuerte efecto ambiental sobre la composición de la comunidad. Es probable que las propiedades intrínsecas de las masas de aguas profundas, además de la presencia de presas o alimento alternativo, permitan la existencia de diferentes nichos tróficos.

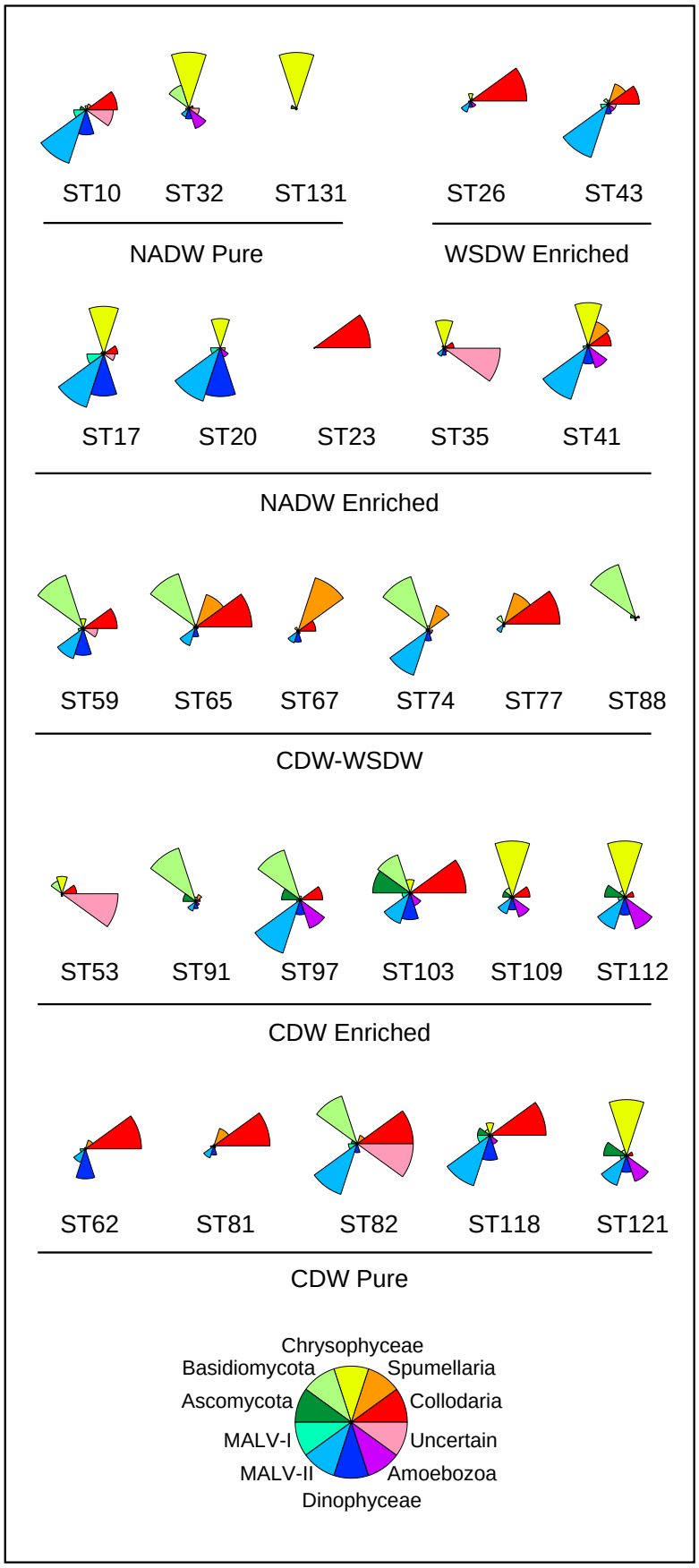


Figura 7, Capítulo 4.La abundancia relativa de los diez grupos filogenéticos más abundantes en todas las muestras profundas. Las estaciones se agrupan por su respectiva masa de agua.

Fagotrofia: la relación con los procariotas

Se esperaba que la fagotrofia fuera la vía trófica principal para los microeucariotas en el océano profundo. Para probar esta hipótesis, primero se analizó la relación entre la abundancia de procariotas y microeucariotas y después la relación entre su ratio y la diversidad. Teniendo en cuenta toda la columna de agua profunda, la abundancia de los microeucariotas se correlaciona bien con la de los procariotas ($R^2 = 0.50$, $p = 0.0001$), excepto en las estaciones del Pacífico Sur ($R^2 = 0.08$, $p > 0.05$, Figura 5, capítulo 3). Posteriormente, los análisis de regresión múltiple demostraron que este resultado era independiente de la profundidad (Capítulo 3). Sin embargo, a pesar de esta relación significativa, los procariotas explican sólo una parte de la varianza de la abundancia de

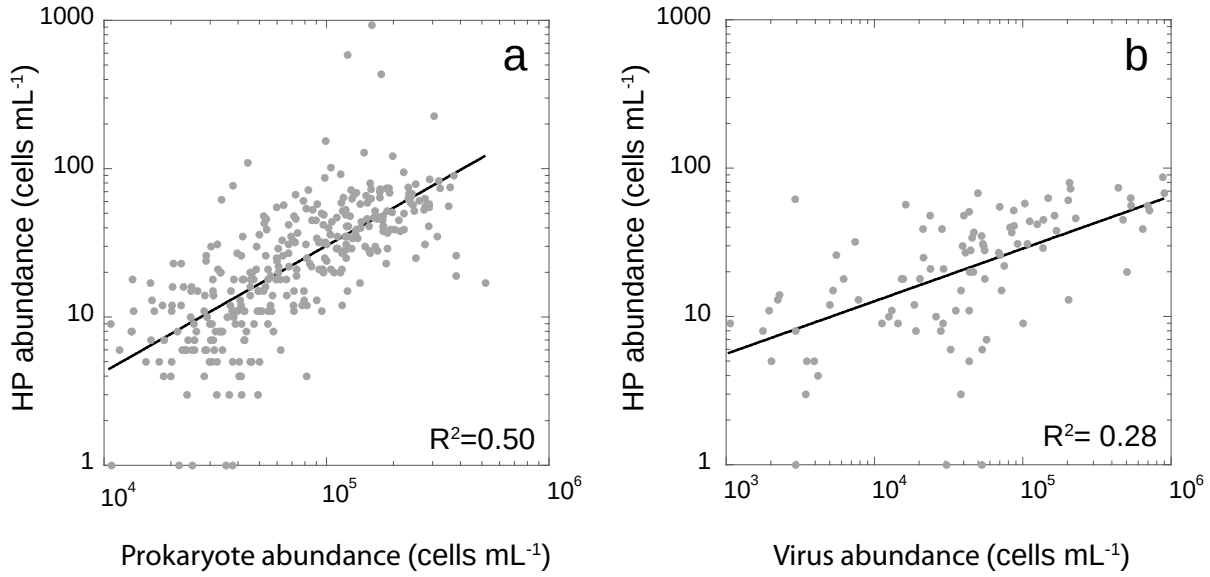


Figura 5, Capítulo 3. Abundancia de los protistas heterotróficos frente a la abundancia de los procariotas (a) y frente a la abundancia de virus grandes (b), en muestras derivadas de 71 y 20 perfiles verticales, respectivamente.

los microeucariotas. La abundancia de los procariotas ha sido utilizada también para explicar la variabilidad en la diversidad, pero el resultado no fue significativo ($p > 0.058$), aunque el 34% de la variabilidad fue explicada significativamente por el ratio entre los procariotas y los microeuca-

riotas ($p = 0.001$). Los valores bajos de esta proporción, similares a los de aguas epipelágicas (ca. 2000), corresponden a comunidades dominadas por Collodaria y Chrysophyceae, lo que sugiere un papel putativo de depredadores para estas dos clases. Teniendo en cuenta que menos de la mitad de la variabilidad de la abundancia y de la diversidad se explica por los procariotas y que el ratio de abundancia tiene un valor promedio más alto que en la superficie, se puede concluir que la fagotrofia en las profundidades del océano parece dar también un espacio importante a las demás vías tróficas tales como la osmotrofia y el parasitismo.

Osmotrofia: el papel de los hongos

Comparando el ratio entre procariotas y microeucariotas con la abundancia relativa de varios taxones, sólo los hongos han presentado una relación significativa (Figura 7, Capítulo 3). Teniendo en cuenta que los hongos son ciertamente osmotrofos y probablemente incapaces de realizar fagotrofia (Richards *et al*. 2012), se podría sugerir que cuando la comunidad estaba dominada por los hongos la presión de depredación sobre los procariotas era menor, favoreciendo así valores altos en el ratio de las abundancias.

Como se ve en la introducción (Figura 7b) la distribución del COD no es constante en la región batipelágica, de hecho el COD disminuye a lo largo de la cinta transportadora profunda, resultando más concentrado en aguas atlánticas que pacíficas. Es posible asociar parte de la disminución del COD desde el Océano Antártico hacia el Pacífico Norte a la presencia de hongos en estas aguas. Sin embargo, teniendo en cuenta que el COD se concentra más en el océano Atlántico, especialmente en el norte, es difícil entender por qué los hongos no prosperan en estas aguas (excepto en la estación 32). Hay varias explicaciones posibles:

- *Hongos versus procariotas*. Una primera hipótesis simple es que los hongos compiten con los

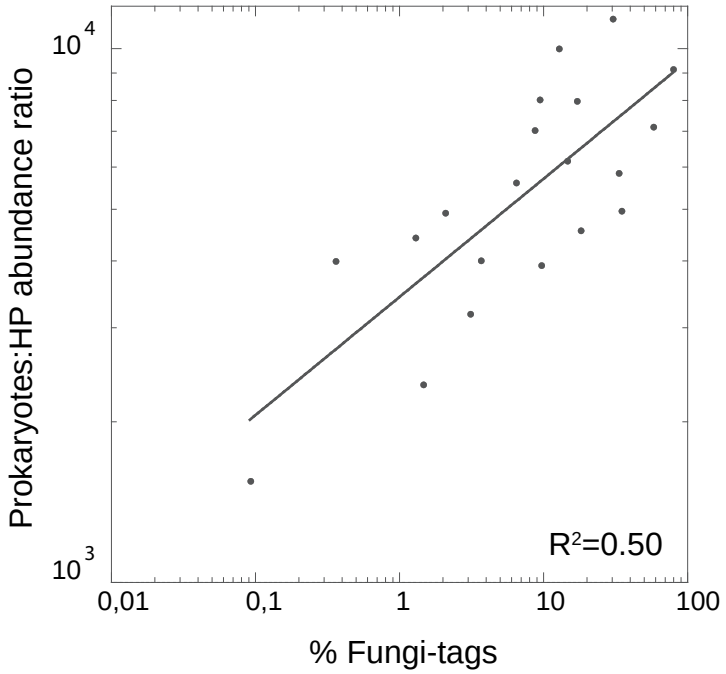


Figura 7, Capítulo 3. Relación entre el ratio de las abundancias de procariotas respecto a las células de HP y el porcentaje de secuencias de hongos en las muestras correspondientes. Los porcentaje derivan de un estudio paralelo sobre la diversidad de protistas de profundidad (Pernice et al, en preparación).

procariotas por el COD en las aguas del Atlántico. Una relación antagonista de los hongos y bacterias se observó en varios experimentos, como por ejemplo en Moller *et al.* (1999), donde los hongos y procariotas compiten claramente por el COD.

- *Hongos versus Crisófitas*. Las Crisófitas pueden ser fagotrofas (Massana 2011) y osmotrofas (Sandgren *et al.* 1995, Sanders *et al*. 2001), y pueden sobrevivir gracias a una combinación de estas dos estrategias. Salvo en la estación 32, nunca comparten el dominio con los hongos. Våge *et al*. (2013) construyeron un modelo con el fin de probar la importancia de la mixotrofia en comparación con la osmotrofia pura. Demostraron que a baja proporción de tamaño entre presas (procariotas) y depredadores (crisófitas), como sucede en las profundidades del océano, para un mixotrofo es muy conveniente la estrategia de "comerse al competidor" (Thingstad *et al.* 1996).

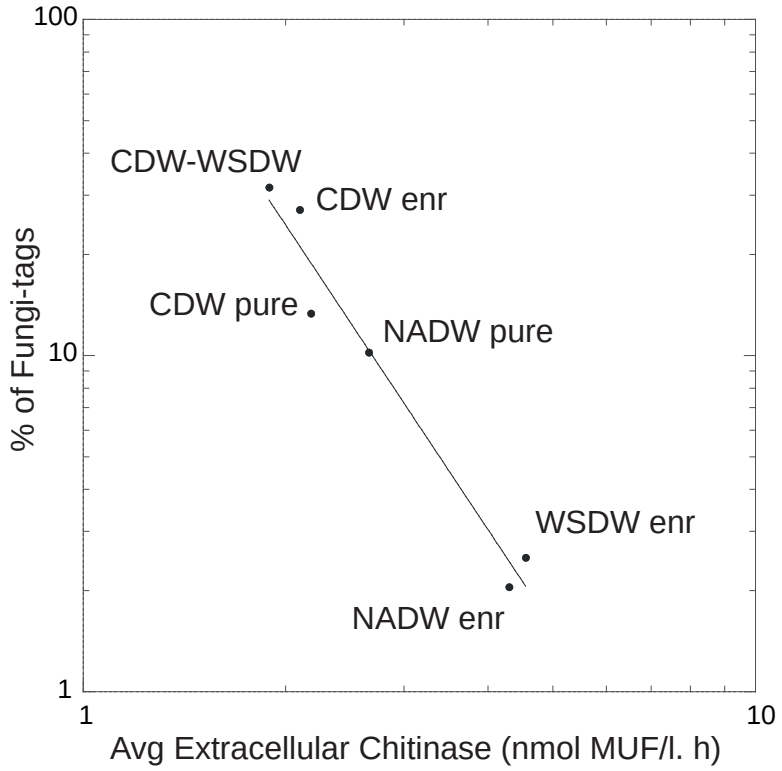


Figura 9. Relación entre la concentración de quitinasa extracelular y la abundancia media de secuencias de hongos en las diferentes masas de agua definidas.

Así que en este caso los mixotrofos (crisófitas) podrían suprimir los osmotrofos puros (hongos) depredándolos, como podría ocurrir en aguas del Atlántico y del Pacífico Norte.

- *Hongos y carbono recalcitrante*. La presencia mayor de hongos en aguas pobres en COD podría explicarse por una especialización a la asimilación de carbono recalcitrante. Un posible mecanismo es la secreción de moléculas de superóxido, en particular, formas oxidadas de Mn, que oxida carbono recalcitrante convirtiéndolo en formas más biodisponibles (Hansel *et al.* 2012). Este mecanismo está probablemente compartido con las bacterias. Por lo tanto, otra explicación de un ratio mayor entre procariotas y microeucariotas es que en una comunidad dominada por Hongos también los procariotas podrían prosperar con el COD recalcitrante.

- *Hongos frente a la quitinasa*. Una de las características típicas de los hongos es la presencia de quitina en la pared exterior (Richard *et al* . 2012). Durante la expedición Malaspina se midió la actividad enzimática extracelular presente en el agua, incluyendo la quitinasa, la enzima que digiere la quitina. La actividad de la quitinasa resulta ser mayor en el Atlántico que en el Pacífico e Índico. Teniendo en cuenta valores promedios para cada masa de agua, la abundancia relativa de las secuencias de Hongos muestran una relación inversa con la actividad de la quitinasa (Figura 9) en una correlación altamente significativa ($p=0.0005$; R^2 de 0.95). Las quitinasas podrían producirse por procariotas así como por microeucariotas (Cottrell *et al*. 2000). No está claro si los hongos son los principales objetivos de esta enzima, de todos modos el efecto de la quitinasa extracelular es un ambiente no favorable para su vida.

La osmotrofia también está presente en otros grupos taxonómicos, como los Labyrinthulidae (Raghukamar *et al*. 2001) o Excavata (Lara *et al*. 2009). De hecho, el alcance del proceso osmotrófico debería estudiarse mejor para entender el impacto de los protistas heterótrofos en el balance global de carbono.

Parasitismo: las relaciones ocultas

Inferir interacciones parasitarias desde los datos de secuenciación es bastante difícil porque no existe una correspondencia clara entre el anfitrión y la abundancia del parásito (Skovgard *et al.* 2014). Varios clados de microeucariotas marinos se consideran sobre todo parásitos, siendo el más abundante el MALV-I y el MALV-II. Numerosas especies dentro de los dinoflagelados y hongos también pueden ser parásitos. Los supuestos anfitriones de todos estos parásitos son otros microeucariotas y también la macrofauna. En nuestra base de datos, la abundancia relativa de las secuencias de MALV-II tiene una correlación significativa con la abundancia relativa de las se-

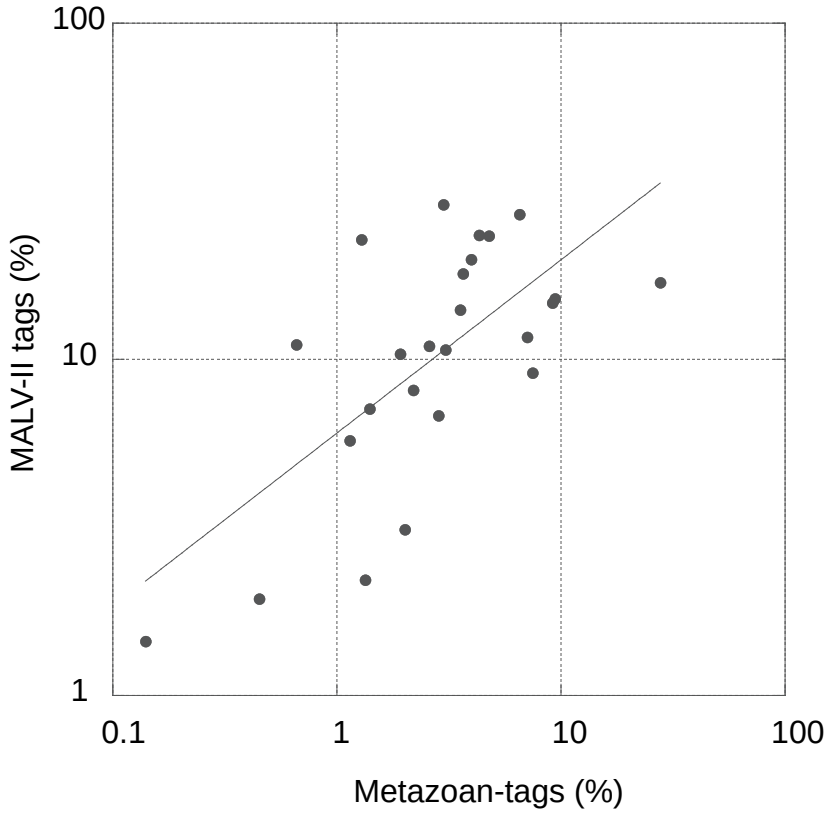


Figura 10. Relación entre la abundancia relativa de la secuencias de MALV-II frente a la de los metazoos.

cuencias de metazoos ($R^2 = 0.45$, $p = 0.0005$) (Figura 10). Esta relación es mejor para MALV-I ($R^2= 0.60$) y menos fuerte para dinoflagelados ($R^2 = 0.41$), pero no se detectó para los hongos u otras clases. Las tres relaciones son particularmente evidentes en las muestras del Atlántico, en la que R^2 es, respectivamente, 0.85, 0.89 y 0.76. Otros candidatos putativos como parásitos del océano profundo son los hongos, especialmente considerando que la OTU principal es muy similar a una especie de parásito reconocida, *Engyodontium álbum*, que parasita *Felis domesticus* (Dennis 1995). Sin embargo, asumiendo una distribución aleatoria de los anfitriones se espera una distribución aleatoria de las secuencias de parásitos. En este sentido, la distribución mejor de secuencias de alveolados encaja con este escenario en comparación con la distribución no aleatoria de hongos.

Las últimas décadas de la investigación de los protistas se han centrado en el estudio de la diversidad de los microeucariotas, al principio definido como "inesperada" en términos de diversidad alta y novedosa. Ahora que esperamos esta increíble complejidad de la composición de los taxones, la atención se está dirigiendo hacia la investigación de la función de los diferentes taxones en el ecosistema. La asignación de un papel ecológico claro, a través del método clásico de cultivo y los nuevos enfoques genómicos de una sola célula, pronto van a mejorar la visión que tenemos ahora sobre este conjunto tan importante del medio ambiente.

Conclusiones

1) La región V4 del gen 18S ADNr representa mejor la variabilidad de todo el gen que la región V9. La pendiente media es de 1.4, este factor podría ser utilizado para obtener la variabilidad de todo el gen.

2) El valor normal de la distancia genética máxima para las secuencias que pertenecen a una misma clase es de 0.25, este valor podría ser útil para evaluar el nivel taxonómico de un ribogrupo.

3) Una típica comunidad epipelágica está constituida por Alveolata (47%), Estramenópilos (19%) y Rhizarian (13%). El resto de la comunidad está compuesto por Archeaplastida y CCTH. A menudo, los Fungi y los Excavata son realmente pocos (menos del 1 %) o no tienen representación.

4) La abundancia de microeucariotas es de 54 ± 5 células mL^{-1} para la capa mesopelágicas y de 14 ± 1 de células mL^{-1} para la batipelágica, su variabilidad se explica principalmente por la profundidad, la abundancia de procariotas y la concentración de oxígeno (en orden de importancia).

5)El tamaño de las células promedio aumenta con la profundidad, el número de células de más de 35 μm^3 (> 4 μm de diámetro) representa el 12% a 200 m y el 22% a 4000 m. La biomasa total varia desde 280 ± 46 pg mL C^{-1} en la capa superior mesopelágica a 50 ± 14 pg mL C^{-1} en la capa más profunda.

6) Los resultados del análisis de la pirosecuenciación 454 se compararon con un análisis metagenómico paralelo. En general, el porcentaje de los supergrupos era muy similar con los dos métodos (Figura 7A). Las diferencias que se encontraron fueron una presencia inferior de Alveolados y una muy superior de Excavata en el metagenoma.

7) Cuatro clases abundantes componen principalmente la comunidad batipelágica: Collodaria, Chrysophyceae MALV-II y Basidiomycota. Mientras que la composición de la comunidad es bastante homogénea entre las muestras la distribución de estas clases es heterogénea.

8) La diferencia en la composición de la comunidad entre las muestras se explica bien debido a la pertenencia a una masa de agua (26 %) y al ratio entre procariotas y microeucariotas (34%).

General references

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS *et al* (2012). The Revised Classification of Eukaryotes. *J Euk Microbiol* **59:** 429-514.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009). A method for studying protistan diversity using massively parrallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4:** e6372, doi:10.1371/journal.pone.0006372.

Amato A, Kooistra W, Ghiron JHL, Mann DG, Proschold T, Montresor M (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* **158:** 193-207.

Arístegui J, Duarte CM, Gasol JM, Alonso-Sáez L (2005). Active mesopelagic prokaryotes support high respiration in the subtropical northeast Atlantic Ocean. *Geophys Res Lett* **32:** L03608, doi:10.1029/2004gl021863

Arístegui J, Duarte CM, Gasol JM, Herndl GJ (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnol Oceanogr* **54:** 1501-1529.

Bass D, Richards T, Matthai L, Marsh V, Cavalier-Smith T (2007). DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* **7:** 162.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V *et al* (2010). Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. *Mol Biol Evolution* **27:** 1698-1709.

Benner R (2002). Chemical composition and reactivity. *In* DA Hansell and CA Carlson (Eds.), *Biogeochemistry of marine dissolved organic matter.* Elsevier Science. pp 59-90.

Boras JA, Montserrat Sala M, Baltar F, Arístegui J, Duarte CM, Vaqué D (2010). Effect of viruses and protists on bacteria in eddies of the Canary Current region (subtropical northeast Atlantic). *Limnol Oceanog* **55:** 885-898.

Burki F, Schalchian-Tabrizi K, Pawlowski J (2008). Phylogenomics reveals a new "megagroup" including most photosynthetic eukaryotes. *Biology Lett* **4:** 366-369.

Burki F, Inagaki Y, Bråte J, Archibald JM, Keeling PJ, Cavalier-Smith T *et al* (2009). Large-Scale Phylogenomic Analyses Reveal That Two Enigmatic Protist Lineages, Telonemia and Centroheliozoa, Are Related to Photosynthetic Chromalveolates. *Genome Biol Evol* **1:** 231-238.

Burki F, Okamoto N, Pombert J-F, Keeling PJ (2012). The evolutionary history of haptophytes and

cryptophytes: phylogenomic evidence for separate origins. *Proc R Soc B* **279:** 2246-2254.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD (2009). Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75:** 5797-5808.

Chambouvet A, Morin P, Marie D, Guillou L (2008). Control of Toxic Marine Dinoflagellate Blooms by Serial Parasitic Killers. *Science* **322:** 1254-1257.

Cheung MK, Au CH, Chu KH, Kwan HS, Wong CK (2010). Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *ISME J* **4:** 1053-1059.

Christaki U, Courties C, Massana R, Catalá P, Lebaron P, Gasol JM *et al* (2011). Optimized routine flow cytometric enumeration of heterotrophic flagellates using SYBR Green I. *Limnol Oceanogr: Methods* **9:** 329-339.

Cottrell MT, Wood DN, Yu L, Kirchman DL (2000). Selected Chitinase Genes in Cultured and Uncultured Marine Bacteria in the α- and γ-Subclasses of the Proteobacteria. *Appl Environ Microbiol* **66:** 1195-1201.

Countway PD, Gast RJ, Dennett MR, Savai P, Rose JM, Caron DA (2007). Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ Microbiol* **9:** 1219-1232.

Díez B, Pedrós-Alió C, Massana R (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environl Microbiol* **67:** 2932-2941.

Dennis, RWG (1995) *Fungi of South East England*. Royal Botanic Gardens, Kew.

Edgcomb V, Kysela D, Teske A, de Vera Gomez A, Sogin M (2002). Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc Natl Acad Sci USA* **99:** 7658 - 7662.

Edgcomb V, Beaudoin D, Gast R, Biddle JF, Teske A (2011). Marine subsurface eukaryotes: the fungal majority. *Environ Microbiol* **13:** 172-183.

Epstein PR, Ford TE, Colwell RR (1993). Health and climate change: Marine ecosystems. *The Lancet* **342**: 1216-1219.

Finlay B (2004). Protist taxonomy: an ecological perspective. *Phil Trans Roy Soc Lond B* **359:** 599 - 610.

Fu Y, O'Kelly C, Sieracki M, Distel DL (2003). Protistan Grazing Analysis by Flow Cytometry Using Prey Labeled by In Vivo Expression of Fluorescent Proteins. *Appl Environ Microbiol* **69:** 6848-6855.

Fukuda H, Sohrin R, Nagata T, Koike I (2007). Size distribution and biomass of nanoflagellates in meso- and bathypelagic layers of the subarctic Pacific. *Aquat Microb Ecol* **46:** 203-207.

Giering SLC, Sanders R, Lampitt RS, Anderson TR, Tamburini C, Boutrif M *et al* (2014). Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507:** 480-483.

Groisillier A, Guillou L, Massana R, Valentin K, Vaulot D (2006). Genetic diversity and habitats of two enigmatic marine alveolate lineages. *Aquat Microb Ecol* **42:** 277-291.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L *et al* (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41:** 597-604.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB *et al* (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *PNAS* **106**:3859-3864.

Hansel CM, Zeiner CA, Santelli CM, Webb SM (2012). Mn(II) oxidation by an ascomycete fungus is linked to superoxide production during asexual reproduction. *PNAS* **109:** 12621-12625.

Hansell DA, Carlson CA (1998). Deep-ocean gradients in the concentration of dissolved organic carbon. *Nature* **395:** 263-266.

Hansell DA, Carlson CA, Repeta DJ, Schlitzer R (2009). Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography* **22**:202–211

Holen DA, Biraas ME (1996) Mixotrophy in chrysophytes. *In*: D Craig, CD Sandgren, JP Smol, J Kristiansen (Eds.) *Chrysophyte algae. Ecology, phylogeny and development*, University Press, Leiden (1996), pp 119–140.

Jebaraj CS, Raghukumar C, Behnke A, Stoeck T (2010). Fungal diversity in oxygen-depleted regions of the Arabian Sea revealed by targeted environmental sequencing combined with cultivation. *FEMS Microbiol Ecol* **71:** 399-412.

Jeon SO, Bunge J, Stoeck T, Barger KJA, Hong SH, Epstein SS (2006). Synthetic statistical approach reveals a high degree of richness of microbial eukaryotes in an anoxic water column. *Appl Environ Microbiol* **72:** 6578-6583.

Jezbera J, Hornák K, Simek K (2005). Food selection by bacterivorous protists: insight from the analysis of the food vacuole content by means of fluorescence in situ hybridization. *FEMS Microbiol Ecol* **52:** 351-363.

Jones RI (2000). Mixotrophy in planktonic protists: an overview. *Freshwat Biol* **45:** 219-226.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12:** 118-123.

Lara E, Moreira D, Vereshchaka A, López-García P (2009). Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environ Microbiol* **11:** 47-55.

Li W (1994). Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol oceanog***:** 169-175.

Libes SM (1992). An introduction to marine biogeochemistry. Wiley

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H *et al* (2013). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* doi: 10.1111/1462-2920.12250

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603 - 607.

Marie D, Simon N, Guillou L, Partensky F, Vaulot D (2001). DNA/RNA Analysis of Phytoplankton by Flow Cytometry. *Current Protocols in Cytometry*. John Wiley & Sons, Inc.

Massana R, Balagué V, Guillou L, Pedrós-Alió C (2004a). Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiol Ecol* **50:** 231-243.

Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A *et al* (2004b). Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol* **70:** 3528-3534.

Massana R, Guillou L, Terrado R, Forn I, Pedros-Alio C (2006). Growth of uncultured heterotrophic flagellates in unamended seawater incubations. *Aquat microb ecol* **45:** 171-180.

Massana R, Pedrós-Alió C (2008). Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11:** 213-218.

Massana R (2009). Pycoeukaryotes. *In*: M Schaechter (Eds.) *Encyclopedia of Microbiology*. Elsevier Science (2009), pp 674–687.

Massana R, Unrein F, Rodríguez-Martínez R, Forn I, Lefort T, Pinhassi J *et al* (2009). Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J* **3:** 588-596.

Massana R (2011). Eukaryotic picoplankton in surface oceans. *Annu Rev Microbiol* **65:** 1-47.

Massana R, Pernice M, Bunge JA, Campo Jd (2011). Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean. *ISME J* **5:** 184-195.

Møller J, Miller M, Kjøller A (1999). Fungal–bacterial interaction on beech leaves: influence on decomposition and dissolved organic carbon quality. *Soil Biol Biochem* **31:** 367-374.

Moon-van der Staay SY, De Wachter R, Vaulot D (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409:** 607-610.

Morgan-Smith D, Herndl GJ, van Aken HM, Bochdansky AB (2011). Abundance of eukaryotic microbes in the deep subtropical North Atlantic. *Aquat Microb Ecol* **65:** 103-115.

Morgan-Smith D, Clouse MA, Herndl GJ, Bochdansky AB (2013). Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. *Deep Sea Res Pt-I* **78:** 58-69.

Nagata T, Tamburini C, Arístegui J, Baltar F, Bochdansky AB, Fonda-Umani S *et al* (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. *Deep Sea Res Pt-II* **57:** 1519-1536.

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Tobe K (2007a). Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315:** 252-254.

Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ (2007b). Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. *Environ Microbiol* **9:** 1233-1252.

Olson RJ, Vaulot D, Chisholm SW (1985). Marine-phytoplankton distributions measured using

shipboard flow-cytometry. *Deep Sea Res* **32:** 1273-1280.

Pernice MC, Logares R, Guillou L, Massana R (2013). General patterns of diversity in major marine microeukaryote lineages. *PLoS ONE* **8:** e57170.

Pernthaler J, Pernthaler A, Amann R (2003). Automated enumeration of groups of marine picoplankton after fluorescence in situ hybridization. *Appl Environ Microbiol* **69:** 2631-2637.

Porter K, Feig Y (1980). The use of DAPI for identifying and counting aquatic microflora. *Limnol oceanog* **25:** 943-948.

Quince (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6:** 639-641.

Raghukumar S, Ramaiah N, Raghukumar C (2001). Dynamics of thraustochytrid protists in the water column of the Arabian Sea. *Aquat Microb Ecoly* **24:** 175-186.

Richards TA, Jones MDM, Leonard G, Bass D (2012). Marine Fungi: Their Ecology and Molecular Diversity. *Ann Rev Mar Sci* **4:** 495-522.

Rosselló-Mora R, Amann R (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25:** 39-67.

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H *et al* (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350-1354.

Salani FS, Arndt H, Hausmann K, Nitsche F, Scheckenbach F (2012). Analysis of the community structure of abyssal kinetoplastids revealed similar communities at larger spatial scales. *ISME J* **6:** 713-723.

Sanders RW (1991). Mixotrophic Protists In Marine and Freshwater Ecosystems. *J Protozool* **38:** 76-81.

Sanders RW, Caron DA, Davidson JM, Dennett MR, Moran DM (2001). *Nutrient Acquisition and Population Growth of a Mixotrophic Alga in Axenic and Bacterized Cultures,* vol. 42.

Sauvadet A-L, Gobet A, Guillou L (2010). Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environ Microbiol* **12:** 2946-2964.

Schlegel M, Meisterfeld R (2003). The species problem in protozoa revisited. *Eur J Protistol* **39:** 349-355.

Schnepf E, Drebes G, Elbrachter M (1990). *Pirsonia guinardiae,* gen. et spec. nov.: a parasitic flagellate on the marine diatom Guinardia flaccida with an unusual mode of food uptake, vol. 44. Biologische Anstalt Helgoland: Hamburg, germany.

Seenivasan R, Sausen N, Medlin LK, Melkonian M (2013). *Picomonas judraskeda* Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as 'Picobiliphytes'. *PLoS ONE* **8:** e59565 doi:10.1371/journal.pone.0059565

Siano R, Alves-de-Souza C, Foulon E, Bendif EM, Simon N, Guillou L *et al* (2011). Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* **8:** 267-278.

Skovgaard A (2014). Dirty Tricks in the Plankton: Diversity and Role of Marine Parasitic Protists. *Acta Protozool* **53**:51-62

Sohrin R, Imazawa M, Fukuda H, Suzuki Y (2010). Full-depth profiles of prokaryotes, heterotrophic nanoflagellates, and ciliates along a transect from the equatorial to the subarctic central Pacific Ocean. *Deep Sea Res Pt-II* **57:** 1537-1550.

Stoeck T, Taylor G, Epstein S (2003). Novel eukaryotes from a permanently anoxic Cariaco Basin (Caribbean Sea). *Appl Environ Microbiol* **69:** 5656 - 5663.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology* **7:** 72.

Tanaka T, Rassoulzadegan F (2002). Full-depth profile (0–2000 m) of bacteria, heterotrophic nanoflagellates and ciliates in the NW Mediterranean Sea: Vertical partitioning of microbial trophic structures. *Deep Sea Res Pt-II* **49:** 2093-2107.

Thingstad TF, Havskum H, Garde K, Riemann B (1996). On the Strategy of "Eating Your Competitor": A Mathematical Analysis of Algal Mixotrophy. *Ecology* **77:** 2108-2118.

Våge S, Castellani M, Giske J, Thingstad TF (2013). Successful strategies in size structured mixotrophic food webs. *Aquat Ecol* **47:** 329-347.

Vaulot D, Eikrem W, Viprey M, Moreau H (2008). The diversity of small eukaryotic phytoplankton (≤3 μm) in marine ecosystems. *FEMS Microbiol Rev* **32:** 795-820.

Venter J, Remington K, Heidelberg J, Halpern A, Rusch D, Eisen J *et al* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

von der Heyden S, Chao E, Cavalier-Smith T (2004). Genetic diversity of goniomonads: an ancient divergence between marine and freshwater species. *Eur J Phycol* **39:** 343 - 350.

Wintzingerode F, Göbel UB, Stackebrandt E (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21:** 213-229.

Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *PNAS* **74:** 5088-5090.

Worden A (2006). Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat microb ecol* **43:** 165-175.

Yamaguchi A, Watanabe Y, Ishida H, Harimoto T, Furusawa K, Suzuki S *et al* (2004). Latitudinal Differences in the Planktonic Biomass and Community Structure Down to the Greater Depths in the Western North Pacific. *J Oceanogr* **60:** 773-787.

Zhu F, Massana R, Not F, Marie D, Vaulot D (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52:** 79-92.

Zubkov M, Tarran G (2008). High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* **455:** 224-226.

Zubkov MV, Burkill PH (2006). Syringe pumped high speed flow cytometry of oceanic phytoplankton. *Cytometry Part A* **69A:** 1010-1019.

Zubkov MV, Burkill PH, Topping JN (2007). Flow cytometric enumeration of DNA-stained oceanic planktonic protists. *J Plankton Res* **29:** 79-86.

Agradecimientos

Agradecimientos

Más de una vez me han llamado afortunado y probablemente lo soy. No es el tipo de fortuna que te hace ganar la lotería, o que te hace encontrar dinero en el suelo, es más de esta suerte que hace que las cosas sean leves y que vayan en la dirección correcta, hacia lo bonito de la vida. Estos años han sido una gran muestra de esta fortuna, un río de regalos, estímulos y experiencias que ni siquiera me imaginaba. Esta fortuna no es una diosa con ojos vendados, esta fortuna tiene caras y nombres y hoy quiero agradecerlos.

Ringrazio prima di tutto i miei genitori per avermi dato la possibilitá di arrivare dove sono adesso, sono due persone bellissime e sono fortunato di essere figlio loro. Un ringraziamento per il sostegno e l'allegria a mia sorella Cristina, e ai miei fratelli Marco e Francesco, vi auguro che i vostri sogni e progetti si realizzino.

Agraeixo al meu "cap" Ramon, per haver-me donat la possibilitat de fer un doctorat, i per haver-me ajudat moltíssim en el desenvolupament d'aquesta tesi. Tot i la normal tensió que comporta la fase final segueixo pensant que ha estat una sort tenir un coordinador amb els peus a terra considerant que el meu cap se'n va als núvols.

Agradezco, a la gran familia ICM, a la que estoy contento y orgulloso de pertenecer. La especial calidad de las personas que se pasean en este edificio sigue sorprendiéndome, he tenido la posibilidad de relacionarme con diferentes micro-clusters sociales, que como en mi tesis a menudo se distinguen por su comportamiento trófico. . .

Agradezco a los de "comemos dentro que es más cómodo, a ser posible en la mesa redonda": Raquel por su soporte y su presencia constante, Montse pel riure i l'especial complicitat, Bea por su maravillosa y gran energía, Eli por cuidar de mis aportes de vitaminas, Ero por las sabrosas castañadas, Irene F. por su macroscópica ayuda al microscopio, Pedro muy presente en los domingos de esta última fase y Juanlu por las charlitas naturalísticas.

Agradezco a los de "comemos al solete que es más chulo y por supuesto carajillo de Baileys": Sara por las charlas desahogantes en el despacho, Guillem por las conversaciones científicas y por las conversaciones absurdas, Elena por nuestras indispensables pausas juntos, Fran por su talento pa' cantar y reírse, Anamari por su alma sureña llena de sol, Eli A. por su sonrisa.

Agradezco a los de "hoy como donde me da la gana y sábado hay barbacoa": Rosana que por suerte se quedó más de seis meses, Roy por ser una cascada de alegría, María d F. que sabe lo mucho que me encanta, Ana G. por su dulzura, Rodrigo por haber sido acogedor conmigo, Sarah Jeanne por su estilo loco que me llena de alegría, Suso por su simpatía y Xavi Leal por su buen humor. Y "los nuevos", que nuevos ya no son: Elisa, Mireia, Fran (hola), Encarna y Caterina que va a llegar lejos. Un gracias especial a "los italianos": Rachele, Miriam, Stefano, Elisabetta, parlare ogni tanto italiano é un sollievo.

Agradezco a los "gourmet del tupper" siempre disponibles a echar una mano, Ramiro que es indispensable, Pablo por su gran simpatía y sus muchos postres, Isabel por la alegría en las fiestas, Marta por ser majísima y Bibiana por su ayuda estadística.

Agradezco a los de "vamos al restaurante", todos los jefes del departamento, importante referencia en estos años en particular: Marta Estrada, Carlos Pedrós, Pep Gasol, Silvia Acinas, Rafel Simó, Celia Marrasé, Montse Sala, Dolors Vaqué, Elisa Berdalet, Esther Garcés, Albert Calbet, Miquel Alcaraz y Enric Saiz. Y aprovecho también para agradecer a las técnicos, indispensables en cada paso, Clara e Vanessa, por toda su ayuda en esta tesis.

Agradezco a los de “Jamón, I miss you” que ya se fueron al otro lado del mar: Clara, Juancho, Arancha, Hugo, Ivo, Pati, Thomas, Javier y Mariona, a la que quiero muchísimo. Ha sido bonito compartir todo este tiempo, pienso mucho en vosotros ahora que ya me toca marchar a mí.

Agradezco a los de “Había un Jamón en el Hespérides”, mis queridos amigos malaspinianos, Lara por su gran corazón, Xiker por nuestro viaje, Víctor por su sentido del humor, Laura y Belén por las cervecitas.

Quedan muchas personas más y espero no olvidarme a nadie, gracias a Lorena, Eva Flo, Estela, Sdena, Isabel(miniMi), Dafne, Sergio, Conchita B., Nuria A., Eva L., Uxue, Carlos D., Jordi G., Elena G. Ariadna y Marieta.

Estos años han sido leves también por la presencia de otros amigos que me han apoyado. Quiero agradecer ante todo a “mis chicas”, mi primera familia en Barcelona, con quien siempre estaré en deuda por haberme acogido: a mi bellíssima Esthel que está siempre presente, a Carol que vale mucho, a Nuria por su espíritu guerrero, a Rosa que sabe bien qué es un doctorado y Angie por su dulzura. Un agradecimiento particular va para mi “familia” actual que se ha tragado todo el estrés final de la tesis, a Gemma por el cariño y la buena vida y a Manuel por la buena vida y el cariño. Agradezco también a “mis chicos” compañeros de numerosos desayunos, meriendas y cenitas: Javi y Quim, Samuel, Enrique, Iñaki, Valentino, Rubén, Lluis y Miguel, hacéis mi vida mucho más divertida e interesante. Gracias también a los Bencinis, Alessia y la Familia Blau que colorean mis mañanas. Un gracias muy muy grande a Irene que me ha dado muchísimo. Un gracias especial a Roberta, que de alguna manera ha sido testigo de todos estos años, eres preciosa en mi vida.

Un grazie agli amici dell’universitá, Emanuele, Daniele, Florenza, Tiziana, Teresa e Pina con cui iniziai a studiare biologia. E non poteva mancare un ringraziamento agli amici di tutta una vita: Olga, Antonino, Ilaria, Emanuela, Valeria e Sabina, grazie per essere stati presenti in questi anni e perdonatemi il poco che mi sono fatto sentire.

Os agradezco a todos *mis fortunas,* gracias por haber paseado por mi vida.

“E quindi uscimmo a riveder le stelle.”

Dante alighieri, Inferno XXXIV, 139

Has vist que bé que he parlat?
Quin discurs tan ben travat.
Quins principis clars i ferms,
dignes d'un home de seny.
Però, un avís per navegants:
Fes-me cas els dies senars,
i els parells, fes com qui sent
que a la platja hi xiula el vent.

(Manel)